DOCUMENT RESUME

ED 052 654	FL 002 384
TITLE	Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for its Investigation, and Practical Applications. 1 July - 30 September 1970.
INSTITUTION	Haskins Labs., New Haven, Conn.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Information Systems Research
REPORT NO	SR-23
PUB DATE	Oct 70
NOTE	211p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$9.87
DESCRIPTORS	Acoustics, *Articulation (Speech), Artificial Speech, Auditory Discrimination, Auditory Perception, Behavioral Science Research, *Laboratory Experiments, *Language Research, Linguistic Performance, Phonemics, Phonetics, Physiology, Psychoacoustics, *Psycholinguistics, *Speech, Speech Clinics, Speech Pathology

ABSTRACT

This report is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. The reports contained in this particular number are state-of-the-art reviews of work central to the Haskins Laboratories' areas of research. The papers included are: (1) "Phonetics: An Overview," (2) "The Perception of Speech," (3) "Physiological Aspects of Articulatory Behavior," (4) "Laryngeal Research in Experimental Phonetics," (5) "Speech Synthesis for Phonetic and Phonological Models," (6) "On Time and Timing in Speech," and (7) "A Study of Prosodic Features." (Author)

ED052654

SR-23 (1970)

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 July - 30 September 1970

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

FL002384

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use.)

ACKNOWLEDGEMENTS

The research reported here was made possible in part by support from the following sources:

Information Systems Branch, Office of Naval Research
Contract N00014-67-A-0129-0001
Req. No. NR 048-225

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Research and Development Division of the Prosthetic and Sensory Aids Service, Veteran Administration
Contract V-1005M-1253

National Institutes of Health
General Research Support Grant FR-5596

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Haskins Laboratories, Inc.
270 Crown Street
New Haven, Conn. 06510

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
N/A

3. REPORT TITLE
Status Report on Speech Research, No. 23, July-September 1970

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Interim Scientific Report

5. AUTHOR(S) (First name, middle initial, last name)
Staff of Haskins Laboratories; Franklin S. Cooper, P.I.

6. REPORT DATE
October 1970

7a. TOTAL NO. OF PAGES
218

7b. NO. OF REFS
750

8a. CONTRACT OR GRANT NO.
ONR Contract N00014-67-A-0129-0001

b. PROJECT NO.
NIDR: Grant DE-01774
c. NICHD: Grant HD-01994
NIH/DRFR: Grant FR-5596
d. VA/PSAS Contract V-1005M-1253

9a. ORIGINATOR'S REPORT NUMBER(S)
SR-23 (1970)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
None

10. DISTRIBUTION STATEMENT
Distribution of this document in unlimited.*

11. SUPPLEMENTARY NOTES
N/A

12. SPONSORING MILITARY ACTIVITY
See No. 8

13. ABSTRACT

This report (for 1 July - 30 September 1970) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. The reports contained in this particular number are state-of-the-art reviews of work central to the Laboratories' areas of research:

Phonetics: An Overview
The Perception of Speech
Physiological Aspects of Articulatory Behavior
Laryngeal Research in Experimental Phonetics
Speech Synthesis for Phonetic and Phonological Models
On Time and Timing in Speech
A Study of Prosodic Features

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6811

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

UNCLASSIFIED
Security Classification

A-31408

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Phonetics						
Perception of speech						
Physiology of speech production						
Laryngeal research						
Speech synthesis						
Timing in speech						
Prosodic features						
Electromyography						

DD FORM 1 NOV 65 1473 (BACK)
S/N 0101-807-6821

CONTENTS

I. Introductory Note 1

II. Extended Reports and Manuscripts

Phonetics: An Overview 3

The Perception of Speech 15

Physiological Aspects of Articulatory Behavior 49

Laryngeal Research in Experimental Phonetics 69

Speech Synthesis for Phonetic and Phonological Models 117

On Time and Timing in Speech 151

A Study of Prosodic Features 179

III. Manuscripts for Publication, Reports, and Oral Papers 209

INTRODUCTORY NOTE ABOUT STATUS REPORT 23

This issue of the Status Report on Speech Research contains state-of-the-art reviews of work central to the Laboratories' areas of research. They were written by staff members for Volume XII of the series Current Trends in Linguistics. This series, edited by Thomas A. Sebeok and published by Mouton & Co., began to appear in 1963. Fourteen volumes are projected, the final volume to appear in 1974. Each takes as its domain linguistic scholarship in a particular geographical area or an aspect of the field of linguistics. Specialists have contributed chapters to each volume assessing the state of knowledge and research activity in areas in which they themselves are productive.

Volume XII is to appear in 1972 in three tomes under the title Linguistics and Adjacent Arts and Sciences. It is to include the following sections:

Part One: History of Linguistics
Part Two: Linguistics and Philosophy
Part Three: Semiotics
Part Four: Linguistics and the Verbal Arts
Part Five: Special Languages
Part Six: Linguistic Aspects of Translation
Part Seven: Linguistics and Psychology
Part Eight: Linguistics and Sociology
Part Nine: Linguistics and Anthropology
Part Ten: Linguistics and Economics
Part Eleven: Linguistics and Education
Part Twelve: Phonetics
Part Thirteen: Bio-Medical Applications
Part Fourteen: Computer Applications
Part Fifteen: Linguistics as a Pilot Science

In addition to Thomas A. Sebeok, the editor of the series, Arthur S. Abramson, Dell Hymes, Herbert Rubenstein, and Edward Stankiewicz are serving as associate editors, and Bernard Spolsky as assistant editor of this volume.

The associate editor in charge of Part Twelve, Arthur S. Abramson, as well as five contributors, Michael Studdert-Kennedy, Katherine S. Harris, Ignatius G. Mattingly, Leigh Lisker, and Philip Lieberman are affiliated with Haskins Laboratories. A sixth contributor to that part, Masayuki Sawashima, did most of his work on his contribution while he was a guest member of the Haskins research staff, although he is normally at the University of Tokyo. Mr. Peter de Ridder of Mouton & Co. has kindly allowed us to follow our normal procedure of including in our Status Reports on Speech Research manuscripts accepted for publication elsewhere. The circumstances are unusual in that this issue of our Status Report comprises a sizeable bloc of material soon to appear under the copyright of a publishing house. We trust Mouton's courtesy to us will be repaid if the sampling of articles presented here stimulates interest on the part of our readership in acquiring the whole volume for their institutions or themselves once it has been released.

A.S. Abramson's "Overview" necessarily refers to all ten articles in his section and not just the five contributions from Haskins Laboratories. For the convenience of our readers, we have listed below the Table of Contents for Part Twelve: Phonetics.

Arthur S. Abramson	Phonetics: An Overview
D.B. Fry	Phonetics in the Twentieth Century
John M. Heinz	Speech Acoustics
Michael Studdert-Kennedy	The Perception of Speech
Katherine S. Harris	Physiological Aspects of Articulatory Behavior
Masayuki Sawashima	Laryngeal Research in Experimental Phonetics
Ignatius G. Mattingly	Speech Synthesis for Phonetic and Phonological Models
Leigh Lisker	On Time and Timing in Speech
Philip Lieberman	A Study of Prosodic Features
J.C. Catford	Phonetic Field Work
André Malécot	Crosslanguage Phonetics

Phonetics: An Overview[*]

Arthur S. Abramson[+]
Haskins Laboratories, New Haven

Phonetics is traditionally concerned with the ways in which the sounds of speech are produced, but the resulting descriptions normally mix auditory factors with articulatory ones, thus depending ultimately upon percepts of the phonetician. This would be true even in a laboratory report using such terms as "high back vowel" that do not themselves stem from instrumental analysis. At the very least, then, we must include the hearing of speech sounds in the definition of phonetics to the extent of allowing for the behavior of the field phonetician who uses his ears to match spans of speech with points or zones of the reference grid he has learned. This grid consists of auditory images correlated with places and manners of articulation. That is, the practical phonetician uses auditory phonetics as a research technique in achieving the goals of articulatory phonetics.

In recent decades, with the waxing importance of psychology in phonetic research, there is no question but that auditory perception has become a central topic of phonetics. In addition, the rise of experimental phonetics with its rapidly improving instrumental techniques (Fant, 1958; Cooper, 1965) has made it possible to look at the speech signal itself, thus adding acoustic phonetics to the scope of the field. In the light of these developments, phonetics may now be defined as the study of the speech signal and its production and perception.[1] A broad view of the interweaving of practical phonetics, the study of the production of speech, analysis of the acoustic signal, and experiments on perception is presented within an historical framework by Dennis B. Fry in his article in this volume.

The collaboration of phoneticians,[2] acousticians, electrical engineers, experimental psychologists, and physiologists has enabled phonetics to surge forward in recent decades, but at the same time it tends to hamper the linguist in applying the findings of phonetic research to his own phonological preoccupations. Even if mild professional indignation prompts one to rebuke the linguist whose phonological abstractions seem to be unsupported by the facts of speech production and perception (Abramson and Lisker, 1970; Fry, this volume), it certainly behooves us phoneticians to present our material from time to time in a form that our linguist colleagues will find readable.[3] It is hard to think of a textbook in phonetics published in the last decade that fills this gap. Some (e.g., Abercrombie, 1967) provide a general theoretical and factual matrix within which to give a course. Others (e.g., Gimson, 1962; Malmberg, 1963; Schubiger, 1970) try to do some justice to the union of linguistic phonetics, acoustics, physiology, and psychology already mentioned. There are also a few textbooks that gently introduce their readers to rather technical material (e.g., Ladefoged, 1962; Denes and Pinson, 1963; Hadding-Koch and Petersson, 1965; Zemlin, 1968; Lindner, 1969); others with this orientation might be considered textbooks, but only by those with some sophistication in mathematics and electronics (e.g., Fant, 1960; Flanagan, 1965). Certain monographs specifically addressed to linguists have been readable enough to be vulnerable to scholarly criticism as well as appreciated for their possible impact on points of linguistic analysis and theory (e.g., Abramson, 1962; Ladefoged, 1964; Delattre, 1966; Lieberman, 1967; Gårding, 1967; Lehiste, 1970).

---

[*] Introduction prepared for Current Trends in Linguistics, Vol. XII. Thomas A. Sebeok, Ed. (The Hague, Mouton).

[+] Also, University of Connecticut, Storrs.

[1] I believe that nowadays it is pointless to insist on a sharp distinction between experimental, or instrumental, phonetics and articulatory-auditory, or "practical," phonetics.

[2] The term here includes scholars who traditionally function in academic departments of linguistics and modern languages as well as those who function, sometimes under the label of "speech scientists," in departments of speech and hearing.

Even the well-motivated linguist, then, cannot have found it convenient to keep abreast of new developments in phonetics over the last several years. Scanning the proceedings of the various international congresses, such as the International Congress of Phonetic Sciences, is a haphazard way of doing this. In the absence of comprehensive textbooks, a satisfactory way of presenting material on the state of the art and science of phonetics is to invite a group of specialists to contribute chapters to a book. Although one such collection has appeared recently, the new edition of the Manual of Phonetics (Malmberg, 1968), it was felt that it would be appropriate and useful for a volume of Current Trends in Linguistics to include another collection of papers with a considerable change in the selection of major topics as well as a largely different list of authors. Two of the authors, J.C. Catford and D.B. Fry, reappear, but their versatility has permitted us to ask them to contribute chapters on new topics.

The foregoing considerations should not lead the reader to believe the collection of papers on phonetics in the present volume can in fact serve as an introductory textbook in phonetics. In accepting a topic, each author committed himself to a critical exposition of the research in areas in which he himself is active. In deciding on a list of ten topics and ten authors to handle them, I tried to think of themes that would cut broad swaths through the field of phonetics in an interesting and useful fashion. Some, of course, are more narrowly specialized than others and consequently more demanding of the reader. The authors, by the way, were not passive partners in the selection of topics; some of them negotiated for deviations from the original agreement. For the linguist dipping into this section then, the kind of background he would need for intelligent reading is available in the books cited earlier. Internally there is considerable cohesion among the ten articles presented here. They might all best be read against the background of the article by D.B. Fry and perhaps some of the thoughts expressed in this "Overview." The authors were aware of each other's presence in the projected volume, some of them even being in the fortunate position of being able to see a few of their fellow contributors' rough drafts, and they provide cross references where needed. In what follows, I will try to furnish more systematic interconnections.

---

[3] Naturally this point applies equally to the grammarians.

DIRECTIONS

Modern phonetics has moved in a number of directions simultaneously, aiming to achieve a greater understanding of the phenomena of the production and perception of speech, making contributions to phonological and psychological theory, and yielding practical advances of use to language teachers and communication engineers. A lucid and well-balanced perspective on these directions and goals is to be found in D.B. Fry's article.

Production and Perception

It is obvious that one can attend separately to mechanisms of production of speech (Harris, Sawashima) and perception of speech (Studdert-Kennedy), but in any broad view of the speech process, it is becoming increasingly difficult to keep the two apart. The study of the correlations between these two aspects of the speech event, with particular concentration on the acoustic signal as the link that binds them, has been called acoustic phonetics. The article contributed by J. Heinz was intended as a statement of what we know today about speech acoustics. In fact, treating the topic under the three divisions of sources of sound, sound modification, and acoustic characteristics of speech sounds, it goes much further. By shuttling between aspects of physiological control and acoustic output, the author provides the basis for a combined articulatory-acoustic phonetics that, taken together with the study of speech perception, I believe to be the kind of phonetics that all students of linguistics should have as part of their training. The reader with such objectives in mind will find it useful to go directly from Heinz to Studdert-Kennedy. Of course, I should add that concern with the relations between production and perception is found throughout this section of the book, with some authors stressing one side or the other.

Phonology

Phonology, whether it be considered an autonomous domain of language structure or a component of the grammar intimately integrated with other components (Postal, 1968), is likely to be fairly abstract. The linguist may need broad phonetic symbols to represent phonemes or matrices of distinctive features, for use in transcribing sentences of a language in a distinctive fashion or spelling lexical entries. But no matter how abstract the level of phonology--a phonemic transcription devoid of all redundancy or sets of binary classificatory features for characterizing underlying forms--the phonologist's strategy must include instructions for going from his particular kind of abstraction to a phonetic output that is consistent with observational data on both production and perception. In seeking to determine the physical bases of phonological distinctions, the phonetician provides information on physiological mechanisms and acoustic features implied by the phonologist's instructions. To what extent the linguist is willing to check his phonological statements against the data supplied by the phonetician (Lisker et al., 1962) may depend in part upon the depth of his conviction as to the validity of his theory and in part upon his level of sophistication in matters of speech production and perception; the two may not be unrelated. Students of language vary considerably in their willingness to let observation and experiment alternate with speculation and theorizing; this is the constant theme that runs through D.B. Fry's article.

While current theorizing on generative phonology is perhaps in many respects more abstract than the kinds of phonology espoused by other schools of thought, it also has the merit of trying to be very explicit about the speech-producing capabilities of the human vocal tract in terms of a "universal phonetics" (Chomsky and Halle, 1968: Chap. 7). Of course the very explicitness of these physical descriptions of phonetic features should, and indeed does, make at least some of them vulnerable to criticism based on hard data or, in fact, on lack of data (Lisker and Abramson, 1971). It should be understood that other phonetic theories have been presented in the recent past (e.g., Peterson and Shoup, 1966a, b), but it is especially encouraging occasionally to find an experimental phonetician, well grounded in linguistics, laying the basis of a phonetic theory within a phonological framework (Lieberman, 1970). The argument that the phonologist, confronted by the increasingly technical nature of phonetics, can avoid yielding to "the temptation to do phonology on the basis of phonetic folklore" by focusing his attention on the work of investigators engaged in synthesis of speech by rule is persuasively presented by I.G. Mattingly. Other articles in this section that would seem to be relevant to particular principles of phonology are those by L. Lisker, P. Lieberman, and J.C. Catford, even though all the contributions should be useful to the linguist in their bearing on general phonetics.

Psychology

It must be stressed that the interdisciplinary character of phonetic research today is not meant merely to provide grist for the linguist's mill. One of the participating disciplines, psychology, also has a vested interest in phonetics. Speech is such a complex, and yet so highly organized, form of human behavior that it was inevitable that psychologists would raise questions about it and design experiments to answer those questions. Here, of course, we are concerned with those psychologists who have investigated mechanisms involved in the production of the sounds of speech as well as the perceptual processing of those sounds. Two of the authors in the present collection, K.S. Harris and M. Studdert-Kennedy, are psychologists who have devoted the major part of their research efforts for some years to phonetics. Most of the other authors have worked in close collaboration with experimental and physiological psychologists.

I think it is easy to defend the proposition that it was the emergence of the Haskins Laboratories group (Cooper, 1950; Cooper et al., 1951) that stimulated lines of research that gradually convinced more and more psychologists that psychologically interesting questions could be answered by experimental phonetic techniques.[4] One of the members of the group, M. Studdert-Kennedy, leads us carefully and revealingly through what might otherwise be a maze of disconnected roads and byways covering recent thought and research into the perception of speech. His generous bibliography enables the reader to pursue with ease any point raised in his account. It may surprise some readers to learn that much physiological phonetic work going on in our day derives from psychological speculation concerning links between certain aspects of speech perception and the control of articulatory gestures. In her chapter in this volume, K.S. Harris discusses the matter with particular attention to electromyography as a research technique. Finally, it must be said that psychologists have begun to find phonetic research relevant to questions of language acquisition in both children and adults.

---

[4] Harvey Fletcher (1929) and George A. Miller (1951) provide surveys of earlier work on speech perception.

METHODS AND APPLICATIONS

In the past decade, the rapidly increasing involvement of engineers, physicians, and computer people in speech research has resulted in a great elaboration of methods of conducting phonetic studies. The authors in this section discuss some of these techniques only to the extent that they are needed in support of themes and concepts being presented. The reader who desires a broad but not too technical knowledge of these developments should consult one of the general works mentioned at the beginning of this "Overview." Anyone wishing to avail himself of a new technique or an instrument not commercially available may find what he needs in such technical notes as occasionally appear in the Journal of the Acoustical Society of America (e.g., Shipp et al. 1970; Sussman and Smith, 1970). Interwoven with references to research methods in the following paragraphs is concern with the more or less practical application of phonetics to problems of communications engineering, speech and hearing disorders, language description, and language pedagogy.

Physiological Research

Physiological phonetics has furnished us with a rather detailed, if uneven, picture of supraglottal articulations, control of the larynx, management of movements and accumulations of air and--in recent years--muscle contractions involved in all these aspects of the speech event. What with current speculation on "feature detectors" and data bearing on the probable links between perception and articulation, it is to be hoped that the neurologists will help us probe into speech mechanisms at even higher levels of control in the not too distant future.

If the linguist is right in asserting that a phonological entity can appear intact over a wide range of environments, then one important task of the phonetician is to explain physiologically what production mechanisms are common to all these manifestations and, at the same time, which ones are needed to account for contextual variation. Of the latter, some may be under active control of the speaker and others simply automatic consequences of the constraints of the human vocal apparatus. With these questions as a unifying theme, K.S. Harris gives us a critical survey of the major trends in current research on speech physiology. After establishing a theoretical framework along with a helpful digression on electromyography, she presents the organization of the speech musculature. This is divided into the respiratory system, laryngeal mechanisms, and the upper articulators. Linguists and, indeed, phoneticians who have uncritically accepted certain phonetic observations as support for particular phonological hypotheses may find it sobering to read some of the discussion in this part of Harris's contribution. The rest of the review concerns the organization of speech, followed by a

12

concluding discussion of the failure so far to find an invariant physiological representation of the phoneme at the peripheral levels investigated.

K.S. Harris gives considerable attention to laryngeal mechanisms as does P. Lieberman in his review of work on prosodic features. Of course there has been considerable physiological research on the larynx itself, to be found mostly in the medical literature. M. Sawashima, one of those rare laryngologists devoting much of their effort to speech research, has culled this literature to provide background material to help the phonetician understand the mechanisms whose functions he is exploring. Following this, the bulk of Sawashima's report discusses recent progress in observing the larynx during the production of voice and speech. This is not an article for beginners. To reap all the benefits to be had from it, the reader should at least have the depth of knowledge of the anatomy of the larynx and of the myoelastic-aerodynamic theory of phonation as described by Sonesson (1968) or Zemlin (1968). P. Lieberman presents physiological data and arguments dealing with the interplay of variations in tensions of the intrinsic muscles of the larynx and changes in subglottal air pressure in the control of prosodic features.

Acoustic Analysis

The availability of the sound spectrograph at the end of World War II (Joos, 1948) gave a great impetus to the research effort that culminated in the present-day acoustic theory of speech production so succinctly outlined by J. Heinz in his report on seminal studies of the last couple of decades. Here, too, we have an article that requires some background on the part of the reader. It would be advisable to have control of such elementary acoustic concepts and basic acoustic phonetics as presented in Denes and Pinson (1963) before reading Heinz to learn about current trends and findings in acoustic phonetic research. The articles by L. Lisker on temporal aspects of speech and A. Malécot on studies in comparative phonetics lean heavily on acoustic data. Of course, most of the other articles make frequent allusions to acoustics too.

Speech Synthesis

Gone are the days when speech synthesizers were available only to the privileged few at scattered points in Europe and North America. Experimental phoneticians now have access to synthesizers at many universities and research institutes. Nearly all of them are terminal analog devices which, when properly programmed, simulate the acoustic output of the human vocal tract. At the same time, there has been work on synthesizers that are analogs of the vocal tract itself. The synthesizer can be used as a very flexible linguistic "informant," capable of separately controlling individual phonetic parameters in a way no human speaker can do. I.G. Mattingly gives a helpful historical and conceptual background on speech synthesizers before launching into the questions of their relevance for phonetic and phonological models. Here, too, it must be said that we are talking of a method of phonetic research that looms large in most of the articles in this section.

Experiments in Speech Perception

Manipulating real or synthetic speech has been a powerful research technique with two major objectives: (1) finding the information-bearing elements of the speech signal and (2) testing hypotheses on the nature of speech perception. For the first goal, the experimenter examines acoustic displays of utterances, usually spectrograms, to pinpoint features that seem to be correlated with the phonological distinction of interest. Nowadays, as compared with the pioneer days of this enterprise, he would approach the task armed with an acoustic theory of speech production that takes articulation into account. Having formed a hypothesis as to what is carrying the information, in the simplest case a single acoustic feature, he will synthesize a set of utterances varying only along this single dimension, record them many times each on magnetic tape and then play them in random order to native speakers of the language for identification as words or syllables of the language. For example, let us suppose that our investigator wishes to determine what acoustic cues distinguish /s/ from /ʃ/. Examining such pairs of English words as sin/shin, so/show, etc., where he believes on linguistic grounds the same phonological contrast to prevail over all the environments, he observes certain frequency differences in the spectral distribution of the turbulent noise of the fricatives.[5] He will then set the parameters of the synthesizer so that appropriate formants and nasal resonances are combined to give the auditory impression of [ɪn] with a time span reserved at the beginning for the frication variants. In the initial slot, he uses the noise generator of the synthesizer to produce variants in small steps over the full range of spectral differences observed and perhaps somewhat beyond to be sure to bracket the two phonemes. The rest of the procedure is as I have outlined it for the general case. Having found that differences in spectral distribution of noise energy are indeed relevant for this syllable type (Harris, 1956), the phonetician might well check it across a sampling of other vocalic environments. For some kinds of acoustic cues it is possible to avoid synthesis and simply manipulate natural speech, as in filtering experiments (Gay, 1970) or tape cutting and splicing (Hadding-Koch and Abramson, 1964).

For the second goal, the testing of hypotheses on the nature of speech perception, a favorite technique through the years has been discrimination testing. One constant theme has been the comparison of the acuity of discrimination of variants along a physical dimension relevant to a phonological distinction with the findings of the classical psycho-acoustic experiments on the discrimination of pitch, loudness, etc. In more recent years, with questions of lateralization of speech processing in the brain coming to the fore, a prominent technique has been that of the dichotic experiment in which competing signals are presented to the two ears. These topics in their proper setting are presented at considerable length by M. Studdert-Kennedy.

Engineering Applications

Much of the impetus for phonetic research during the twentieth century has come from outside linguistics. Communication engineers concerned with more efficient transmission of speech signals and automatic recognition of speech have contributed much to our understanding of phonetic phenomena

---

[5] Of course he may detect other differences as well, some of which may ultimately turn out to have perceptual relevance too. Normal practice would be to creep up on these one by one, testing the sufficiency of each one, and only later to assess the combined effect of all of them.

(Cherry, 1957; Flanagan, 1965). The early efforts of the telephone engineers concentrated on the problem of getting a sufficiently broad frequency range out of their equipment to cover enough of the voice range for minimum loss of message intelligibility (Fletcher, 1929). Gradually the orientation of these engineers shifted to the analysis of the speech signal into its information-bearing components and the question of what kind of channel was needed to transmit only the features relevant for message intelligibility (Cherry, 1957). One of the fond hopes of our engineering colleagues has been to succeed so well in determining the acoustic cues of speech that it would be possible to design various devices that could "recognize" speech automatically (Flanagan, 1965:158-164). One could then give dictation to "phonetic" typewriters, sort packages in the post office by naming the destination aloud, run machinery by voice command while having the hands free to cope with other aspects of the work, and in general "converse" with computers. Despite much frustration among workers in this field, perhaps largely because of naivete with regard to the syntactic and semantic aspects of speech communication, the work on automatic recognition has helped in our general research effort. D.B. Fry devotes a section of his article to phonetics and engineering.

Handicaps in Communication

Speech therapists and audiologists are ready consumers of phonetic data (Halpern, ms.). It is easy to see that the therapist seeking to help his patient compensate for organic deficiencies or adjust to a post-operative state should be well grounded in phonetics. The same reasoning applies, of course, to the speech correctionist working with normally endowed people whose articulatory habits are for one reason or another abnormal. Many a linguist, however, may not be aware of the applications of phonetic research to the handling of hearing impairments (Whetnall and Fry, 1964; Huntington et al., 1968). Taking a patient's hearing loss into account, the problem, broadly speaking, is to make sufficient acoustic cues available to such residual hearing as he has. Such considerations are important for the design of sensory aids such as conventional hearing aids or more sophisticated devices that may be available in the future. Reading machines for the blind form another class of sensory aids depending very heavily on phonetic research (Cooper et al., 1969). The goal here is to have a machine that will scan the printed page and, using speech synthesis by rule (Mattingly, this volume), convert the printed material into a phonetic output that is not only intelligible but also esthetically quite tolerable to the blind. Much of the acoustic phonetic research conducted at Haskins Laboratories over the years has been applied to this problem.

Practical Phonetics

Linguists, language teachers, and speech therapists are often called upon to apply auditory-articulatory techniques to the description of speech signals. The linguist does it as part of his code-cracking operation in doing field work with an unknown language. The language teacher does it to assess the errors of his students. The speech therapist does it in the clinic or classroom to describe deviations from normal speech. How well and consistently can a practical phonetician describe a vowel phone using, say, the IPA Cardinal Vowel reference system? What significance do we attach to such descriptions as "a slightly backed [y]" as compared with "a slightly fronted [ʉ]"? Although it is true that for some linguistic purposes a "phonetic" transcription is desired that reflects the intuitions of the native speaker (Chomsky and Halle, 1968:14), it is also important as part of our account of speech behavior to have narrow transcriptions of utterances as uninfluenced as possible by linguistic bias. For those of us who accept the latter argument, it is at the same time necessary to be sensitive to the problem of calibrating the practical phonetician's ability to take his own percepts of stretches of speech, segment them into phones, and describe these speech sounds usefully in terms of their production and auditory quality (Ladefoged, 1960; Witting, 1962; Laver, 1965). A comprehensive discussion of these matters is provided by J.C. Catford. Some relevant thoughts are also expressed by D.B. Fry in the section of his article that deals with linguistic phonetics. The reader should also consult L. Lisker's contribution, particularly for problems of segmentation and length.

Language Teaching

For the language teacher, the typical phonological account is much too superficial. I put it this way purposely even though it may disturb the linguist to think of a good phonetic description as reflecting anything more than a rather superficial stratum. For the teacher wrestling with the problem of making his students overcome the phonic interference of their native language and master the articulatory patterns of a foreign language, a somewhat better phonetic description is required than is generally found in the linguistic literature. By now many contrastive studies of groups of languages aimed at such a goal are available. The phonetic rules incorporated in their phonological analyses normally derive from articulatory-auditory techniques (e.g., Moulton, 1962). Obviously there is much room for application of instrumental phonetics to these pedagogical problems. It is perhaps not surprising that the impact of this kind of speech research on language teaching does not as yet appear to have been very great. For example, in the early 1960's F.S. Cooper and I, in collaboration with various linguists, produced a set of X-ray motion pictures in slow motion with stretched speech. These films of supraglottal articulations in English, Hungarian, Mandarin Chinese, Russian, and Syrian Arabic were sponsored by the United States Office of Education, not for use in the language classroom itself but for the training of language teachers in the phonetics of these several languages. Although individual specialists in these languages have used the films for their own research purposes, it is not evident to me that departments of language teaching in schools of education have been eager to make much use of them.[6] In recent years a few experimental phoneticians have devoted more of their time and energy to questions of comparative phonetic analysis yielding data that should be useful in language teaching. Since some of these people are accepted on other grounds as members of the confraternity of language teaching methodologists, their work may have a greater impact. The report by A. Malécot surveys most of what has been done in this field and serves as a guide to the relevant literature.

---

[6] I prefer to believe that this is through disinclination rather than dissatisfaction with the quality of the films.

CONCLUSION

The choice of authors and topics, as well as the organization of this "Overview," reflects my own outlook and that of close colleagues. This should not be taken to mean that nothing else of interest has been done or should be done in phonetic research or in the application of such research to other areas. For example, some scholars may wish to examine poetry (Fónagy, 1961) and other types of literature with phonetic features in mind. Others have shown how the methods of experimental phonetics can be applied to research on children's acquisition of speech (Eimas et al., 1971). Among linguists, the historical phonologist might be well advised in positing formulaic representations of protolanguages to be more concerned with phonetic plausibility than he often is.

It is to be hoped that the ten reports on the state of the art and science of phonetics presented in this volume will stimulate much interest on the part of linguists. I readily admit that the list of these distinguished workers in the field could easily have been extended to include a number of others, but this is always so in such a collection.

REFERENCES

Abercrombie, David. 1967. Elements of general phonetics. Chicago: Aldine.

Abramson, Arthur S. 1962. The vowels and tones of standard Thai: Acoustical measurements and experiments. Bloomington: Indiana U. Res. Center in Anthro., Folklore and Linguistics, Pub. 20.

Abramson, Arthur S. and Leigh Lisker. 1970. Laryngeal behavior, the speech signal and phonological simplicity. To appear in the Proceedings of the Tenth International Congress of Linguistics, Bucharest, 1967, Vol. IV.

Cherry, Colin. 1957. On human communication: A review, a survey, and a criticism. Cambridge, Mass.: M.I.T.

Chomsky, Noam and Morris Halle. 1968. The sound pattern of English. New York: Harper.

Cooper, Franklin S. 1950. Spectrum analysis. JAcS. 22.761-2.

Cooper, Franklin S. 1965. Instrumental methods for research in phonetics. Proceedings of the Fifth International Congress of Phonetic Sciences, ed. by Eberhard Zwirner and Wolfgang Bethge, 142-71. Basel: Karger.

Cooper, Franklin S., Alvin M. Liberman and John M. Borst. 1951. The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proc. Nat. Acad. Sci. 37.318-25.

Cooper, Franklin S., Jane H. Gaitenby, Ignatius G. Mattingly and Noriko Umeda. 1969. Reading aids for the blind: A special case of machine-to-man communication. IEEE Trans. Audio. 17.266-70.

Delattre, Pierre. 1966. Studies in French and comparative phonetics. s'Gravenhage: Mouton.

Denes, Peter B. and Elliot N. Pinson. 1963. The speech chain: The physics and biology of spoken language. Murray Hill, N.J.: Bell Telephone Laboratories.

Eimas, Peter D., Einar R. Siqueland, Peter Jusczyk, and James Vigorito. 1971. Speech perception in infants. Science 171.303-6.

Fant, C. Gunnar M. 1958. Modern instruments and methods for acoustic studies of speech. Proceedings of the Eighth International Congress of Linguists, ed. by Eva Sivertsen, 282-358. Oslo: Oslo Univ. Press.

Fant, Gunnar. 1960. Acoustic theory of speech production. The Hague: Mouton.

Flanagan, James L. 1965. Speech analysis, synthesis and perception. Berlin: Springer.

Fletcher, Harvey. 1929. Speech and hearing. New York: Van Nostrand.

Fónagy, Iván. 1961. Communication in poetry. Word 17.194-218.

Gårding, Eva. 1967. Internal juncture in Swedish. Lund: Gleerup. (Trav. de l'Institut de Phonétique de Lund, VI.).

Gay, Thomas. 1970. Effects of filtering and vowel environment on consonant perception. JAcS. 48.993-8.

Gimson, A.C. 1962. Introduction to the pronunciation of English. London: Edward Arnold.

Hadding-Koch, Kerstin and Arthur S. Abramson. 1964. Duration versus spectrum in Swedish vowels: Some perceptual experiments. Studia Linguistica 18.94-107.

Hadding-Koch, Kerstin and Lennart Petersson. 1965. Instrumentell Fonetik: En Handledning. Lund: Gleerup.

Halpern, Harvey (ed.) ms. Communicative disorders: An introduction to speech pathology, audiology and speech science. New York: Random House (to appear).

Harris, Katherine S. 1956. Some acoustic cues for the fricative consonants. JAcS. 28.160-1 (Abstract).

Huntington, Dorothy A., Katherine S. Harris, and George N. Sholes. 1968. An electromyographic study of consonant articulation in hearing-impaired and normal speakers. JSHR. 11.149-58.

Joos, Martin. 1948. Acoustic phonetics. Baltimore: Linguistic Soc. Am. (Also, Suppl. to Lg. 24, no. 2, April-June 1948).

Ladefoged, Peter. 1960. The value of phonetic statements. Lg. 36.387-96.

Ladefoged, Peter. 1962. Elements of acoustic phonetics. Chicago: U. of Chicago.

Ladefoged, Peter. 1964. A phonetic study of West African languages: An auditory-instrumental survey. Cambridge: Cambridge U. Press.

Laver, J.D.M.H. 1965. Variability in vowel perception. Language and Speech 8.95-121.

Lehiste, Ilse. 1970. Suprasegmentals. Cambridge, Mass.: M.I.T.

Lieberman, Philip. 1967. Intonation, perception, and language. Cambridge, Mass.: M.I.T.

Lieberman, Philip. 1970. Towards a unified phonetic theory. Linguistic Inquiry 1.307-22.

Lindner, Gerhart. 1969. Einführung in die experimentelle Phonetik. München: Max Hueber.

Lisker, Leigh, Franklin S. Cooper and Alvin M. Liberman. 1962. The uses of experiment in language description. Word 18.82-106.

Lisker, Leigh and Arthur S. Abramson. 1971. Distinctive features and laryngeal control. To appear in Lg. 47.4.

Malmberg, Bertil. 1963. Phonetics. New York: Dover.

Malmberg, Bertil (ed.) 1968. Manual of phonetics. Amsterdam: North Holland.

Miller, George A. 1951. Language and communication. New York: McGraw-Hill.

Moulton, William G. 1962. The sounds of English and German. Chicago: U. of Chicago Press.

Peterson, Gordon E. and June E. Shoup. 1966a. A physiological theory of phonetics. JSHR. 9.5-67.

Peterson, Gordon E. and June E. Shoup. 1966b. The elements of an acoustic phonetic theory. JSHR. 9.68-99.

Postal, Paul M. 1968. Aspects of phonological theory. New York: Harper and Row.

Schubiger, Maria. 1970. Einführung in die Phonetik. (Sammlung Göschen Band 1217/1217a). Berlin: Walter de Gruyter.

Shipp, Thomas, Barbara V. Fishman, Philip Morrissey, and Robert E. McGlone. 1970. Method and control of laryngeal EMG electrode placement in man. JAcS. 48.429-30.

Sonesson, Bertil. 1968. The functional anatomy of the speech organs. In Manual of phonetics, ed. by Bertil Malmberg, 45-75. Amsterdam: North Holland.

Sussman, Harvey M. and Karl V. Smith. 1970. Transducer for measuring lip movements during speech. JAcS. 48.858-60.

Whetnall, Edith and D.B. Fry. 1964. The deaf child. London: Wm. Heinemann.

Witting, Claes. 1962. On the auditory phonetics of connected speech. Word 18.221-48.

Zemlin, Willard R. 1968. Speech and hearing science: Anatomy and physiology. Englewood Cliffs, N.J.: Prentice-Hall.

The Perception of Speech[*]

Michael Studdert-Kennedy[+]
Haskins Laboratories, New Haven

> "In short, for the study of speech sounds on any level whatsoever, their linguistic function is decisive." Jakobson, Fant, and Halle (1963:11)

INTRODUCTION

To perceive speech is to extract a message, coded according to the rules of a natural language, from an acoustic signal. The signal is a more or less continuously varying acoustic wave, usually generated by the speech organs of a human or by some device (such as a telephone, synthesizer, or mynah bird) that has been constrained to imitate human speech. The message is a string of lexical and grammatical items that may be transcribed as an appropriately marked sequence of discrete phonemic symbols. How analog signal is transformed to digital message has been studied intensively for no more than forty or fifty years.

Early work was largely guided by the demands of telephonic communication: its aim was to estimate, for example, how much distortion by frequency bandwidth compression, channel noise, or amplitude peak-clipping could be imposed on the speech wave without seriously reducing its intelligibility. Knowledge accumulated during this period has been reviewed by Licklider and Miller (1951) and by Miller (1951). Recent related work concerned with the effects of noise on speech intelligibility has been reviewed by Webster (1969). Among general conclusions that may be drawn, three are of particular interest to the present discussion. First, the frequency band contributing most to the intelligibility of speech is balanced around 1900 Hz and comprises the four or five octaves between about 200 Hz and 4000 Hz, the region of greatest sensitivity in the human auditory threshold curve. This hints at a not unexpected, biologically determined match between speech signal and perception that research has only recently begun to explore (e.g., Stevens, in press; Lieberman, 1970). Second, speech is highly resistant to distortion: even infinite peak clipping (which reduces the wave to no more than a pattern of zero-crossings) may have surprisingly small effects on intelligibility. Evidently, the speech signal is highly redundant, and the human listener is an adept of impletion, able to supply from within information that is lacking without. Again, only recently has research begun to track down the sources of the listener's information in his linguistic capacity.

---

[*] Chapter prepared for Current Trends in Linguistics, Vol. XII, Thomas A. Sebeok, Ed. (The Hague: Mouton).

[+] Also, Graduate Center and Queens College, City University of New York.

A third conclusion of this early work (as peak clipping studies suggest) is that the key perceptual dimensions are not those of the waveform (amplitude and time) but those of its time-varying Fourier transform, as displayed in a spectrogram (frequency, intensity, time). Development of the sound spectrograph (Koenig et al., 1946), followed by publication of a work on "Visible Speech" by Potter et al. (1947) and of a monograph on "Acoustic Phonetics" by Joos (1948), paved the way for the first important task of any perceptual study: definition of the stimulus. In this undertaking, research has been increasingly guided by developments in linguistic theory concerning the structure of the message.

STAGES OF THE PERCEPTUAL PROCESS

Before considering this research, it will be useful to lay out, for purposes of exposition, a rough model of the transformation from signal to message. The process entails, conceptually, at least these stages of analysis: 1) auditory, 2) phonetic, 3) phonological, 4) lexical, syntactic, and semantic. The stages form a conceptual hierarchy but in a working model must be both successive and simultaneous: tentative results from higher levels feed back to lower levels, not only to permit correction of earlier decisions in light of later contradictions but also to permit partial determination of phonetic shape by phonological, syntactic, and semantic rules and decisions.

Only the first stage is based directly on the physical input.[1] It is automatic, that is, beyond voluntary control; it transforms the acoustic waveform that impinges on the ear to some time-varying pattern of neurological events of which the spectrogram is, at present, our closest symbolic representation. What further transformation the signal may undergo is an area of active study (see, for example, Mattingly, this volume; Stevens 1967, 1968a, in press). The process requires at least partially independent, neurological systems for extraction of spectral structure, fundamental frequency, intensity, and duration. These interact and give rise to the auditory (psychological) dimensions of quality (timbre), pitch, loudness, and length. Whether acoustic-psychological transformation already involves neural mechanisms peculiar to speech we will consider below.

All stages beyond the first are abstract: they entail recognition of properties that do not inhere in the signal. Together with our knowledge of social context, they represent the set of expectations, some learned, some probably innate, by which we can (and without which we could not) perceive the signal as speech, speech as language. Training may separate the stages to some degree, and we may demonstrate their psychological reality, initially inferred by linguistic analysis, in the laboratory. But in normal conversation, we are unaware of their contributions to what we perceive. No fixed weights can be assigned to the stages. Indeed, their weights undoubtedly vary with the conditions of listening. Phonetic perception in a high wind is governed as much by situational and higher linguistic factors as by the acoustic signal.

Nonetheless, phonetic perception, the second stage, does bear a peculiarly intimate relation to the first. Though we may, without too much difficulty,

---

[1] See Fry (1956) for this observation and for a discussion of stages in which a slightly different position than the one adopted here is taken.

attach phonetic properties to nonspeech, denying speech its phonetic properties is not so easy. There is a directness in perception that makes it difficult to hear the sounds, even of a totally foreign language, as purely auditory events. We hear them, instead, phonetically. That is to say, we hear them as sounds generated by the vocal organs of a human. The nature of this intermediate, phonetic representation is not known. For, despite the term "phonetic," this level is no longer one of sound but rather of some intricate, abstract derivative from the initial auditory analysis. Perhaps it is not without import that, in struggling to interpret the sounds of an unfamiliar language, we (like children who watch a parent or like ourselves straining for meaning through a glass door) often seek order by articulatory imitation: extraction of phonetic information may be closely tied to production mechanisms. Certainly, traditional phonetics, by its easy switches among terms that purport to describe the sounds and terms that unequivocally describe their presumed antecedents in production, hints at some peculiar relation between audition and articulation. But for the moment, we shelve the question. We take the output of this stage to be isomorphic, with a narrow phonetic transcription or, following the formulation of Chomsky and Halle (1968), with a phonetic matrix: the columns, headed by phonetic symbols, segment the phonetic features of the rows. By this point, segments and categories are already present: the sounds have become speech, if not language. But the message is still redundant and much allophonic variation remains to be resolved.

The third stage of the conceptual hierarchy is phonological: phonetic segment is converted to systematic phoneme (Chomsky, 1966). The stage corresponds to the lowest level of Chomsky and Halle's (1968) generative phonological component, a level at which, according to Chomsky and Miller (1963:fn9), the "phonemes" appear. During this stage, the listener applies phonological rules and determines the status of the perceived "sound" sequence within (or without) his language: he "hears through" the features of the phonetic matrix to the distinctive features of the underlying phonemic matrix. Phonetic analysis will have established, for example, the nasalized medial vowel of [k æ̃ t] in many American English dialects. It remains for phonological analysis to reallocate the nasality from the phonetic column for [æ] to a new column for a following segment and so to arrive at recognition of /k æ nt/"can't" (Malécot, 1956). In this stage, also, the listener may dismiss phonetic information that serves no distinctive purpose in his language, treating, in English, both the initial stop of [t^{h}ap] and the unreleased final stop of [p^{h}at] as instances of /t/. And in this stage, he applies the phonotactic rules of his language to derive an acceptable interpretation of the phonetic information.

We can separate the level experimentally from higher levels by calling for perceptual judgments of nonsense syllables. In cross-language studies, listeners reflect the phonological categories of their native languages by their classification of phonetic segments (Lotz et al., 1960; Abramson and Lisker, 1965; Chistovich et al., 1966; Stevens et al., 1969; Lisker and Abramson 1970). Operation of phonotactic rules within speakers of a single language has also been demonstrated (Brown and Hildum, 1956; Greenberg and Jenkins, 1964). We should note, incidentally, that, for the untrained listener, relations between phonologic and phonetic levels are as close as between phonetic and auditory: under normal conditions, he instantly hears speech according to the phonological categories of his native language.

The fourth and last stage (lexical, syntactic, and semantic) represents a complex of interrelated processes that we wish to exclude from the main line of argument. But there are several points to be made. First, we distinguish between direct and indirect perceptual effects of this stage. By direct effects, we intend those grounded in observable acoustic parameters of the signal. For example, lexical items are marked by the acoustic correlates of stress: variations in duration, intensity, and fundamental frequency (Fry, 1955; Hadding-Koch, 1961; Bolinger, 1958; Fry, 1968). If, in difficult listening conditions, segmental features are partly lost, stress patterns may help delimit the sampling space of the lexical items (Skinner, 1936; Savin, 1963; Kozhevnikov and Chistovich, 1965:238ff.) At the same time, perceived stress does not depend on its acoustic correlates alone. Listeners to a synthetically modified utterance with all vowels displaced to [a], but with timing, frequency, and amplitude variations retained, do a poor job of judging its stress pattern (Lieberman, 1965). They require syntactic or, as Klatt (1969) has argued in discussing Lieberman's results, at least segmental information to make reliable stress judgments (and vice versa: a neat instance of parallel processing). Similar results have been obtained for fundamental frequency contours conveying intonation patterns (Lieberman, 1965). In short, acoustic correlates of higher-order linguistic structures may be directly perceived, although relations between signal and message are relatively loose and we have, as yet, no detailed account of the underlying, interactive process.[2]

More important to the present discussion are the indirect effects on perception of this fourth stage. Here, we intend those provisional syntactic and semantic decisions that may resolve phonetic doubt. As we shall see, even in citation forms, there is frequently no simple, invariant correspondence between spectral structure and phonetic shape; in continuous speech the lack of invariance is even more marked (Shearme and Holmes, 1962). Pickett and Pollack (1963) and Lieberman (1963) found that words excised from sentences and presented to listeners in isolation, without syntactic and semantic context, were poorly recognized. Other studies have shown that words are perceived more accurately in a sentence than on a list (Miller et al., 1951) and have separated the contributions of syntax and meaning to the perceptual outcome (Miller and Isard, 1963).[3]

One may wonder whether these effects of higher-level factors are truly perceptual. A recent study speaks to this question. Warren (1970) demonstrated an effect that he termed "phonemic restoration." Listeners heard a tape-recorded sentence: "The state governors met with their respective legislatures convening in the capital city," with a 120 msec segment deleted and replaced by a cough (and, on another occasion, by a burst of 1000 Hz tone) of the same duration. The missing segment corresponded to the first "s" of

---

[2] Lehiste (1970) reviews the experimental phonetic literature on prosodic features and evaluates its meaning for linguistic theory.

[3] There is also a sizeable literature examining the effects of surface structure on perceptual segmentation of an utterance (Fodor and Bever, 1965; Garrett et al., 1966; Johnson, 1966; Reber and Anderson, 1970) and of deep structure on immediate recall of sentences (Mehler, 1963; Miller, 1964; Savin and Perchonock, 1965; Blumenthal, 1967).

"legislatures," "together with portions of the adjacent phonemes which might provide transitional cues to the missing sound." Of twenty subjects listening to this sentence, nineteen reported that all speech sounds were present, and one reported a missing phoneme but the wrong one. Factors above the phonological may contribute to this illusion: Sherman (cited by Warren, 1970) "found that when a short cough was followed immediately by the sounds corresponding to 'ite,' so that the word fragment could have been derived from several words, such as 'kite' or 'bite,' the listener used other words in the sentence to determine the phonemic restoration; when the preceding and following context indicated that the incomplete word was a verb referring to the activity of snarling dogs, the ambiguous fragment was perceived quite clearly as either 'bite' or 'fight'" (Warren, 1970:393).

The important point is that "the ambiguous fragment was perceived quite clearly": the effect is an illusion[4] and resists manipulation. Warren remarks that other listeners, "despite knowledge of the actual stimulus, still perceived the missing phoneme as distinctly as the clearly pronounced sounds actually present" (p. 392). The study not only demonstrates the abstract nature of phonetic perception, but it is also consistent with "a somewhat novel theory of speech perception" (Chomsky and Miller, 1963:311). This "novel theory" has been summarized by Chomsky and Halle (1968:24):

> The hearer makes use of certain cues and certain expectations to determine the syntactic structure and semantic content of an utterance. Given a hypothesis as to its syntactic structure...he uses the phonological principles that he controls to determine a phonetic shape. The hypothesis will then be accepted if it is not too radically at variance with the acoustic material....Given acceptance of such a hypothesis, what the hearer "hears" is what is internally generated by the rules. That is, he will "hear" the phonetic shape postulated by the syntactic structure and the internalized rules.

This account[5] circumvents the failure of the acoustic signal to meet the "linearity condition" and the "invariance condition" of segmentation into fixed phonetic units (Chomsky and Miller, 1963), problems to which we turn in the following section. However, an adequate theoretical account must not only define the "certain cues" that the hearer uses but also explain how he makes use of them. Like the speaker of Jakobson and Halle (1956:5,6) who, if context and syntax cannot be trusted to take up the slack of slovenly speech, may deploy the full resources of his code to produce "an explicit form which...is apprehended by the listener in all its explicitness," so the listener, unaided by syntax or context, may deploy the code at his command to extract from an intrinsically slovenly acoustic signal an explicit phonetic message. Most of what follows is directed toward an understanding of this act of "primary recognition" (Fry, 1956).

---

[4] Skinner (1936) and Miller (1956) describe similar illusions.

[5] There are counterparts in other areas of perceptuon. Neisser (1966) and Kolers (1968a) review similar problems in visual pattern recognition that have also invited abstract, "constructive" theories of perception.

DEFINITION OF THE STIMULUS

The first task of any perceptual study is to define the stimulus. For this, the sound spectrograph has been the principle instrument, and two complementary methods have been used: analysis and synthesis. Spectrographic measurements first provide data concerning the cues that seem likely to be important for perception (formant patterns, temporal relations, noise bandwidths, and so on). Synthesis is then used to verify or adjust the preliminary conclusions of analysis.

Synthesis was first used in this way at Haskins Laboratories in New York. Cooper (1950) (see also Borst, 1956) developed the Pattern Playback as a research tool for reconverting spectrographic patterns into sound. The patterns, painted on a moving acetate belt, reflect frequency-modulated light to a photo-electric cell that drives a speaker. Portions of the pattern can be systematically emphasized, pruned, deleted until minimal cues for the perception of a particular utterance have been determined (Liberman, 1957). With this device, and with its electronic successors at Haskins and elsewhere, a body of knowledge has been built up concerning acoustic cues for speech, sufficient for synthesis by rule of relatively high-quality speech (Liberman, 1957; Fant, 1960, 1968; Mattingly, 1966, this volume; Flanagan, 1965; Stevens and House, 1970).

Implications of this knowledge have been considered by Liberman et al. (1967b). They draw two pertinent conclusions. First, there are, for the most part, no segments in the acoustic signal that correspond to perceived segments of the message. Certainly the sound stream may be segmented and, as we shall see, these segments may be crucial to perceptual reconstruction of the message. But, whatever level of message unit we look for--distinctive feature, phoneme, or syllable--we frequently find no single sound segment corresponding to it and it alone. There are exceptions: fricatives or stressed vowels, for example, may be isolated in slow speech. But, in general, a single segment of sound contains information concerning several neighboring segments of the message, and a single segment of the message may draw information from several neighboring segments of the sound (see also Fant, 1962, 1968). Since perceptual segmentation not only occurs but is essential to the "duality of patterning" on which human language rests (Hockett, 1958), lack of one-to-one relations between signal segments and message segments constitutes a problem for both psychologists and linguists.

A second, closely related, conclusion of Liberman et al. (1967b) is that acoustically distinct signals (separated by differences that in nonspeech would be well above threshold) are frequently perceived as identical, while acoustically identical signals are frequently perceived as distinct. There is thus, for speech, a lack of isomorphism between sign and percept analogous to that in other areas of perception.

Here, we distinguish between two types of anisomorphism in the perceptual process: one can be observed by the unaided human listener, the other cannot. The first includes "extrinsic" allophonic variations peculiar to a particular language or dialect (Wang and Fillmore, 1961; Ladefoged, 1966) and constitutes a problem in the relations between phonetic and phonological segments (stage 3 of the model outlined above). Thus, in the formulation of Chomsky and Halle (1968), neither columns nor rows of the distinctive feature matrix that serves as input to the generative phonological component are necessarily isomorphic with those of the phonetic feature matrix at output. The inputs /rayt + r/ and /rayd + r/, for example, (see Chomsky and Miller, 1963) emerge as [rayDr] and [ra.yDr]. Here, columnar segmentation of the phonemic input has been lost by transformation of a distinction in the fourth column (voiced/unvoiced) into a distinction in the second (long/short vowel); also, phonologically distinct segments, /t/ and /d/, have become phonetically identical as an alveolar flap, [D], while phonetically distinct diphthongs, [ay] and [a.y], have emerged from the single phonological diphthong /ay/. These transformations may be generated by the ordered application of two phonetic rules. And, in general, the system underlying extrinsic allophonic variations may be inferred and stated in a set of rules relating phonetic and morphophonemic levels. In the present discussion, we shall not further consider this type of variation.

The second type of anisomorphism [and the one to which Liberman et al. (1967) address themselves] was discovered only when it became possible to substitute a suitable analyzing instrument (the sound spectrograph) for the human listener. There then appeared the discrepancies between acoustic signal and phonetic percept referred to above: a lack of one-to-one correspondence between acoustic and phonetic segments and a host of "intrinsic" allophonic variations, such that the acoustic signal clearly could not meet the linearity and invariance conditions imposed by traditional concepts of speech as a sequence of discrete phonetic segments. Anisomorphism of this type constitutes a problem in the relations between acoustic pattern and phonetic segments (stages 1 and/or 2 of the model outlined above).

Attention to these discrepancies was first drawn by perceptual experiments with synthetic speech. For example, Liberman et al.(1952) showed that a brief burst of energy centered at 1440 Hz and followed by a two-formant vowel pattern was sufficient cue for perception of [p] if the vowel was [i] or [u], of [k] if the vowel was [a]. In other words, perception of [p] or [k] was determined not by the frequency position of the stop burst but by the relation of this position to the following vowel. Here, a single acoustic cue controlled two distinct percepts. The authors concluded that, for these stops, "the irreducible acoustic stimulus is the sound pattern corresponding to the consonant vowel syllable." Schatz (1954) confirmed their results in a tape-cutting experiment with natural speech.

Later experiments demonstrated the importance of second formant transitions as cues for distinguishing among labial, alveolar, velar stop, and nasal consonants (Liberman et al., 1954) and went on to show that a sufficient acoustic cue for a given phonetic segment may prove not only different in different contexts but so different that there seems to be no physical basis for perceptual generalization between the tokens (Delattre et al., 1955). In fact, if sufficient cues for [d] in different phonetic contexts (a rapidly rising F_2 transition for [di], a rapidly falling transition for [du]) are synthesized in isolation, their perceptual identity is lost, and they are heard as different "chirps," one rising in pitch, the other falling (Mattingly et al., in press). These and other examples from perceptual studies with synthetic speech are reviewed by Liberman (1957) and by Liberman et al. (1967b).

Studies of natural speech have confirmed that there is enormous variability in the acoustic correlates of a given phonetic segment as a function of phonetic context, stress, and speaking rate (Shearme and Holmes, 1962; Lindblom, 1963; Stevens and House, 1963; Kozhevnikov and Chistovich, 1965; Stevens et al., 1966; Öhman, 1966; Menon et al., 1969). Öhman (1966), for example, collected data from spectrograms. He traced the paths of the first three formants in spectrograms of intervocalic [g], followed and preceded by all possible combinations of five Swedish vowels. He found large variations in the formant transitions on either side of the [g] occlusion as a function of the vowel on the opposite side of the stop. He concluded that "the perception of the intervocalic stop must be based on an auditory analysis of the entire VCV pattern rather than on any constant formant-frequency cue" (p. 167). Thus Öhman implies, as do Liberman et al. (1952), that we reduce acoustic variance to phonetic invariance by analyzing relations between portions of the auditory pattern over sections of roughly syllable length. This process is examined in a later section (Syllables, Segments, Features).

One other invariance problem deserves mention, if only because it fits neither of the types (extrinsic, intrinsic) discussed above: that arising from speaker-dependent variations in vowel formant frequencies. Center frequencies of the first two or three formants as principal acoustic determinants of vowel color have been known for many years (Delattre, et al., 1952; Peterson, 1952, 1959, 1961; Peterson and Barney, 1952; Ladefoged, 1967). Also known since the early work of Fant (1947, reported in Fant 1966) and of Peterson and Barney (1952) is that formant frequencies vary widely enough to produce considerable acoustic overlap between phonetically distinct vowels spoken by different classes of speakers (men, women, children) and by different individuals within a class. Formant frequencies of a vowel spoken by a women tend to be some 10-20 percent higher than those of a phonetically identical vowel spoken by a man, and for children the shift is even greater. What are the grounds of the perceived identity?

The hypothesis that vowels are "normalized" at some point during initial auditory analysis by application of a simple scale factor, inversely proportional to vocal tract length, is probably not tenable (Peterson, 1961). Fant (1966) has brought the problem into relief by showing that the male-female shift for Swedish and American English vowels is not constant for all formant frequency regions (largely due to a greater ratio of male to female tract length in the pharynx than in the mouth cavity), so that the putative normalization factor would differ for rounded back, open unrounded, and close front vowels. Fujisaki and Nakamaura (1969) have developed an algorithm that separates the five vowels of Japanese (spoken by men, women, and children) with more than 90 percent accuracy on the basis of their first two formants,[6] but it is not known whether their method could resolve a richer system or the severely reduced vowels of running speech (Lindblom, 1963). Under these conditions, invariance may be derived at a later stage of the perceptual process.

In fact, evidence exists that listeners adjust their phonetic decision criteria to the speaker's vocal tract characteristics (inferred from F_0 and F_3, involving stage 2 of the perceptual process) (Kasuya et al., 1967; Fujisaki and Kawashima, 1968; Fourcin, 1968) and/or to his vowel quadrilateral

---

[6] See also the work of Pols et al. (1969) on Dutch vowels.

(inferred with aid of situational and linguistic constraints, involving stages 3 and 4) (Joos, 1948; Ladefoged and Broadbent, 1957; Ladefoged, 1967; Gerstman, 1968). However, no solution is yet generally agreed upon, and speaker-dependent vowel variation therefore takes its place beside other invariance problems previously mentioned.

We conclude, then, that while study of the speech signal and its perception has led to an understanding of acoustic cues sufficient for rather successful speech synthesis, it has also revealed that the signal is an intricate pattern of highly variable overlapping acoustic segments, anisomorphic with the perceived message. It has thus raised more problems for speech perception than it has solved. We may bring these problems into focus if we briefly turn our attention to production and ask how "intrinsic" acoustic variability may be presumed to arise.

Over the past dozen years, there has been a number of studies of muscular activity in speech production. Initial impetus for much of the work was given by the notion of the Haskins group that an invariance lacking in the acoustic correlates of phonetic segments might be found among their articulatory correlates. Early electromyographic (EMG) studies (e.g., Harris et al., 1962; MacNeilage, 1963; Harris et al., 1965) offered some support for this hypothesis, but more recent work has not (e.g., Fromkin, 1966; Tatham and Morton, 1968; MacNeilage and DeClerk, 1969), and MacNeilage (1970:184) has remarked: "Paradoxically, the main result of the attempt to demonstrate invariance at the EMG level has been...to demonstrate the ubiquity of variability."[7]

The presumed invariance must therefore lie at some neurological level, presently inaccessible. Some theorists (e.g., Ladefoged, 1966; Öhman, 1965, 1966; Öhman et al., 1967; Liberman et al., 1967) have taken this to be the level of "motor commands" and have assumed muscular variability (and consequent "intrinsic" allophonic variations) to arise from mechanical constraints, neuromuscular inertia, and temporal overlap of successive commands. Others (e.g., Fromkin, 1966; MacNeilage, 1970) have suggested that variability may result from controlled, contextually adapted responses to a set of invariant "go-to" or target commands. Both approaches have been elaborated. Öhman (1967) has developed a mathematical model by which vocal tract shape, area function, and speech wave may be computed from a linear combination of fixed commands and coarticulation functions. MacNeilage (1970) has explored the possibility of controlled, variable responses to fixed target commands in light of current neurological theory. Both describe derivation of a more or less continuous, variable signal from discrete, invariant commands.

If, now, we take these commands to be isomorphic with, if not identical to, some articulatory phonetic feature matrix, such as that proposed by Chomsky and Halle (1968) (cf., Stevens and Halle, 1967; Chistovich et al., 1968; Mattingly, this volume), we have, at least, some way of conceptualizing the nature of the phonetic matrix and of its output relation to the acoustic signal. The problem for perceptual theory is that it has, at present, no firm grip on the reverse relation between acoustic signal and perceptual phonetic matrix. Among the reasons for this are, first, that the processes relating these levels are even less accessible to observation that the corresponding processes on the production side; second, that these levels are so tightly connected perceptually that it is difficult to separate them in behavior; third, that we have no clear concept of, and no terminology to describe, the phonetic matrix at the output of stage 2. Our task is, therefore, to define this abstract, phonetic matrix and its relation to auditory parameters of the acoustic signal.

---

[7] This should not be taken to imply that production is unruly. As MacNeilage's paper makes clear, EMG studies are advancing our understanding of its laws. See also Harris (this volume).

CLASSIFYING SPEECH SOUNDS

Let us begin by considering how we classify speech sounds Experimental evidence reinforces our intuitive recognition that we do so rapidly and involuntarily (see, for example, Kozhevnikov and Chistovich 1965:222 ff.). In this, speech is sui generis. Walking through the woods, we instantly recognize the sound of a waterfall, but with little difficulty, we may choose to hear it as a senseless rumble. Similarly, we have little difficulty in suspending our cognition of a speaker's meaning. But we do find it difficult not to recognize his sounds as speech: recognition is automatic, instantaneous. In what follows, we attempt to disentangle auditory from phonetic (and phonological) stages and to estimate their roles in perception.

Listeners can certainly make purely auditory judgments of speech signals. Flanagan (1965: Ch. VII) has reviewed studies carried out with the general intent of setting upper and lower limits on the discriminability of acoustic dimensions known to be important in speech (vowel formant frequencies, formant amplitudes, formant bandwidths, fundamental frequency, fricative noise bandwidth, and so on). The results, where comparable, give values of the same order as those reported in nonspeech auditory psychophysical studies. But for this, two conditions are necessary: first, the signals must be relatively sustained; second, the listener must be instructed either explicitly or implicitly, by the nature of the experimental task, to listen to the signals as though they were not speech. If the signal is presented in a word or phrase, or among a set of phonetically opposed sounds, a distinctive speech mode of response tends to appear.

For example, Lane et al. (1961) (see also Lane, 1965) asked subjects to judge the loudness of the vowel [a], produced in isolation, and determined a loudness function exponent of 0.7, a value close to that usually found in experiments with nonspeech sounds. Ladefoged and his colleagues (see Ladefoged, 1967:35-41, for a summary of their work) asked subjects to assess the relative loudness of two words in a constant carrier sentence. They found a loudness function exponent of 1.2. This was exactly equal to the exponent of their function relating loudness to the rate of work done upon air in phonation. Ladefoged and his colleagues concluded that their results reflected a distinctive speech mode by which loudness of sound is judged in terms of the physiological effort required to produce it. Lehiste and Peterson (1959) reached a similar conclusion in a study of the loudness of a set of nine steady-state vowels.

Analogous results have been reported in studies of intonation contours. Hadding-Koch and Studdert-Kennedy (1963, 1964) varied the extent and direction of the terminal glide of a fundamental frequency contour imposed synthetically on a vocoded carrier word. They asked subjects, under one experimental condition, to judge the glide as either rising or falling, under another condition, to judge the word as a question or statement. Subjects' psychophysical judgments were influenced by their linguistic judgments: they tend to judge falling glides of words they considered questions as rising, and rising glides of words they considered statements as falling. In an extension and replication of this study (Studdert-Kennedy and Hadding, in preparation) the authors compared psychophysical judgments of contours imposed on a word with those of matched modulated sine-waves. The previously observed effects were much reduced in the sine-wave judgments. Lieberman (1967), in a theoretical account of these results, argues that listeners perceive intonation contours in terms of the subglottal pressure changes and laryngeal maneuvers required to produce them.

In short, if speech sounds are isolated and of fairly long duration, listeners will make reliable auditory judgments of the same order as they make for comparable nonspeech sounds. But if signals are presented in a context that encourages the listener to deploy his linguistic resources, a characteristic mode of perception appears: unable to separate auditory from phonetic, the listener bases supposedly auditory judgments on phonetic or linguistic decisions. By the same token, if experimental conditions permit auditory judgment, listeners may supplement phonetic skills with auditory, provided the signal is of sufficiently long duration. This appears to be a principal basis of differences observed in the discrimination of consonants and vowels.

A typical experiment goes as follows. Two or more phonetic segments are selected for study. Among reasons for selecting the segments is that they are distinguished by acoustic differences lying along a continuum, such as direction of a formant transition, duration of a silent interval, or center frequencies of formants. One of the selected segments is synthesized, usually within a nonsense word or syllable, sometimes (if a fricative or vowel) in isolation. The relevant acoustic cue is then varied systematically, in steps large enough to be psychophysically detectable in nonspeech, small enough for there to be several steps within phonetic categories. The result is a series of a dozen or so acoustic patterns that range in phonetic type from, say, [ba] through [da] to [ga], or from [ta] to [da], or from [i] through [I] to [ɛ]. Several tokens of each pattern are recorded and gathered into random test orders.

Listeners are then asked to identify and to discriminate between acoustic patterns. Identification is usually by forced choice: listeners assign each token to one of the designated phonetic categories.[8] For discrimination, listeners usually hear tokens in triads, of which two are identical, the other different by one or more steps along the continuum, and are asked either to pick out the odd one or to indicate whether the third token is the same as the first or second.

Under these conditions a listener's performance typically approaches one or other of two ideal modes of perception, termed "categorical" and "continuous"

---

[8] As a matter of fact, experiments customarily prescribed phonological categories and, therefore, engage phonological perception. But since our interest here is to separate auditory from phonetic, we disregard the phonological component in what follows.

(Liberman et al., 1957; Liberman et al., 1961b; Fry et al., 1962; Eimas, 1962; Studdert-Kennedy et al., 1970b). By "categorical[9] perception" is intended a mode in which each acoustic pattern, whatever its context, is always and only perceived as a token of a particular phonetic type. Asked to discriminate between two acoustic patterns, the listener can do so if he assigns them to different phonetic categories but not if he assigns them to the same phonetic category. In other words, he can find no auditory basis for discrimination and so must rely on category assignments. By "continuous perception" is intended a mode in which a listener may, if asked, group different patterns into a single category, but his categories are not clearcut (due to context effects), and he is still able to discriminate between patterns that he assigns to the same category. In other words, discrimination is independent of category assignment. (For a fuller account, see Studdert-Kennedy et al., 1970).

Listeners have approached these two modes of perception in many studies. In general, they tend to perceive a continuum categorically if the acoustic variations separate stop consonant categories (e.g., [b, d, g], [p, b], [t, d]), continuously if identical variations are carried by nonspeech signals with no phonetic significance[10] or if the acoustic variations separate sustained vowels.[11]

The categorical/continuous distinction between speech and nonspeech is fundamental. But the same distinction between consonants and vowels is more troublesome. Early interpretations (e.g., Fry et al., 1962; Liberman, et al., 1967a) proposed two distinct perceptual mechanisms: a motor reference mechanism for the categorical stop consonants, paralleling their articulatory discontinuities, an auditory mechanism for the continuously graded vowels. There are many reasons why this is not satisfactory, not least, the difficulty of believing that the syllable, an articulatory and perceptual integer, compounded of consonant and vowel, is analyzed by two distinct mechanisms. Furthermore, this account has been superseded.

Recent work has demonstrated that continuous perception of vowels is a function of their duration and of the experimental method used to study them. Stevens (1968b) has shown that medial vowels in CVC syllables are more categorically perceived than the same vowels sustained in isolation. Fujisaki and Kawashima (1969) have shown that listeners' reliance on category assignment for discrimination increases as the duration of synthetic vowels is reduced from 6 to 3 to 1 glottal pulse. Sachs (1969) has demonstrated a similar effect of duration for vowels in isolation and in word context.

Fujisaki and Kawashima (1969) have also developed a quantitative model of the listener's behavior in discrimination studies. Briefly, the model states that the degree of categorical perception depends on whether auditory or phonetic short-term memory is summoned for the decision process during discrimination. The reliability of auditory short-term memory is less for the brief acoustic events that signal stop consonants than for the relatively sustained events that signal steady-state vowels. The listener has recourse to auditory discrimination whenever he is asked to distinguish between two identical phonetic types; in this, his vowel auditory memory serves him better than his consonant auditory memory, and his vowel discrimination is accordingly superior. The model makes quantitative predictions that have been repeatedly confirmed in experimental tests. The authors conclude that vowels may be perceived either categorically or continuously depending on experimental conditions.

---

[9] The term "categorical" is here preferred to "categorial," since it carries, in addition to the meaning "of or involving a category," shared by both words, the sense "absolute, unqualified." (Webster's Third New International Dictionary, 1965).

[10] See Liberman et al., 1957; Liberman et al., 1961 a,b; Bastian et al., 1961; Eimas, 1963; Fujisaki and Kawashima, 1969; Abramson and Lisker, 1970; Mattingly et al., in press.

[11] See Abramson, 1961; Fry et al., 1962; Eimas, 1963; Stevens et al., 1969.

Chistovich and her colleagues reached the same conclusion by a different route (Chistovich et al., 1966a, 1968). Chistovich has explicitly addressed herself to problems of phonetic classification. She has questioned the value of studying speech discrimination on grounds that the procedure invites a listener to search the signal for auditory qualities that he would not normally detect and so to hear it as nonspeech (1968:34, 35). She has confined her own studies to absolute identification, asking subjects to shadow, mimic, or transcribe natural or synthetic speech sounds. We will not describe the methods here (see below: Syllables, Segments, Features). But in an important paper (Chistovich et al., 1966a; see also Fant, 1968) she has demonstrated that even isolated, steady-state vowels may be perceived categorically, if the experimental method forces the listener's attention to phonetic, rather than auditory, qualities.

We should not be misled into supposing that there are no important differences between consonants and vowels: their functional opposition within the syllable is fundamental to both production and perception of speech. Vowels are acoustically more variable and phonetically more subject to the effects of context than are consonants; they carry a lighter segmental load and virtually all the auditory load of prosodic and indexical features.[12] Their perceptual passage therefore leaves an auditory residue that the listener may put to nonphonetic use: his judgments are then continuous. If the residue is reduced by rapid speech, or if the listener's attention is diverted from it by some resolutely phonetic task, his judgments are categorical. In short, he may perceive vowels both auditorily and phonetically; consonants he perceives phonetically. The distinction to be drawn is not, therefore, between consonants and vowels, but between continuous auditory and categorical phonetic perception, the first typical of nonspeech, the second peculiar to speech.

The argument of the last few pages has brought us no closer to separating the auditory from the phonetic stages of speech perception. But it does suggest that, insofar as listeners can achieve this separation, they are judging auditory aspects of the signal irrelevant to phonetic perception, and that, insofar as they perceive phonetically, they cannot achieve the separation. In other words, the output of stage 1 cannot be brought into consciousness under normal listening conditions.[13] We might even suppose (although this is not a necessary conclusion) that phonetic classification of speech already begins during the initial auditory analysis of stage 1.

---

[12] Abercrombie (1967: Ch. 1) distinguishes between linguistic and indexical (dialectal, personal) features of the acoustic signal.

Such a position is implicit in the "immanent approach" of distinctive feature theory. Theorists emphasize that correlates of the features are to be found at every level of the speech process (articulatory, acoustic, auditory) and that the invariance to be sought in the signal is "relational" rather than absolute (Jakobson et al., 1963; Jakobson and Halle, 1956; Chomsky and Miller 1963). They stress the perceptual importance of the entire spectral pattern rather than of band-limited cues, such as a single formant transition (Fant, 1964). The relational concept is difficult, since reference points for spectral relations must vary with speaker, dialect, phonetic context, stress pattern, and speaking rate. Further, Fant has remarked that "statements of the acoustic correlates to distinctive features have been condensed to an extent where they retain merely a generalized abstraction insufficient as a basis for the quantitative operations needed for practical applications" (Fant, 1962), and no one has attempted to use the acoustic specifications of distinctive features to synthesize speech.

Stevens, in his recent work (1967, 1968a, in press) undertakes to remedy this situation by showing that there is "some justification on a purely physical basis for a characterization of phonemes in terms of discrete properties or features" (Stevens, in press). His general procedure is to compute from an idealized vocal tract model the spectral poles and zeros associated with, for example, a particular point of closure or constriction. For certain points of constriction, there appears a significant concentration of spectral energy; the frequency position of this concentration proves relatively insensitive to small shifts in position of the constriction. These "quantal places of articulation...are optimal from the point of view of sound generation" (Stevens, 1968:200), since they permit relatively imprecise articulation without serious perturbation of the signal. Furthermore, they tend to correspond to places of articulation used in many languages (e.g., velar, postalveolar (retroflex), postdental). Similar computations for [i, a, u], the pivots of most vowel systems, provide spectral correlates of their distinctive feature definitions, in terms of F1-F2-F3 positions (Stevens, in press).

We note, incidentially, that Chistovich et al., (1966b) have reported related perceptual data for vowels. They used a handmanoeuvered version of the Stockholm Royal Institute of Technology's OVE I b. The instrument permits a subject to trace any selected path through the F1-F2 plane (with F_o, F3, and F4 set at appropriate values) and to judge the resulting sounds. Subjects indicated whenever the continuously changing vowel crossed a boundary into a region of altered phonetic quality. Over a hundred such boundary points were determined by four subjects, marked on an F1-F2 plot and connected by best-fitting straight lines. On this plot most boundaries were either vertical (fixed F1) or horizontal (fixed F2), suggesting that "extremely simple rules employing critical boundary values of formant frequencies operate in vowel perception" (Chistovich et al., 1966b; see also Fant, 1968).

---

[13] Day (1968, 1969, 1970) and Day and Cutting (1970a, b) report the results of work with dichotic and other experimental methods that may serve to separate the stages behaviorally.

Implicit in Stevens' work is the assumption that there should prove to be a biologically comfortable match between articulatory and auditory capacities (cf., Halle, 1964; Stevens and Halle, 1967; Stevens et al., 1969). Lieberman (1970) has developed this position more fully, arguing that phonological features may have been selected through a combination of articulatory constraints and "best matches" to specific neural acoustic detectors. Recent work in neurophysiology has demonstrated the existence of relatively complex property, or feature, detectors in cat (Whitfield and Evans, 1965) and frog (Frishkopf and Goldstein, 1963; Capranica, 1965). It is not unreasonable to suppose that comparable detectors, tuned to features of speech, may exist in man.

In short, there are arguments and some evidence to suggest that linguistically relevant features of the acoustic signal may be extracted during initial auditory analysis. How fixed pattern detectors could resolve intrinsic allophonic variations and the incipient entropy of running speech is not clear. But let us suppose that sets of property detectors are, indeed, neatly sprung by the flow of speech. There would then remain the deeper task of grouping the outputs of these detectors into phonetic segments. For this, more than an auditory analysis is required.

SYLLABLES, SEGMENTS, FEATURES

Problems of segmentation have bedeviled speech research since its inception.[14] Here, particularly, research has relied on linguistic theory for definition of the perceptual terminus and has sought to validate postulated theoretical units empirically. In this, students have had the support of linguists. Jakobson and Halle, for example, emphasized the "immanent approach which locates the distinctive features and their bundles within the speech sounds, be it on their motor, acoustical, or auditory level" (1956:8). More recently, Halle (1964) has implied that universal phonetic features may be grounded in man's innate auditory capacities, while Chomsky and Halle take phonetic features to be "identical with the set of properties that can in principle be controlled in speech" (1968:295) and assume each feature to have (presumably discoverable) "acoustical and perceptual correlates" (1968:299).

We are not, however, entirely at the mercy of theory, nor even of possibly illusory perception, in our choice of perceptual units. If we are willing to make the assumption that perceptual units are isomorphic with production units, we have in errors of speech a natural body of materials from which to infer segments. Any unit subject to errors of metathesis (Spoonerism), substitution, or omission must be under some degree of independent control in production. Fromkin (1970) has analyzed six hundred errors collected by herself and her colleagues over three years. The observed units of error pertinent to this discussion were: syllables (e.g., "butterpillar and catterfly"), phone-length segments (e.g., "the nipper is zarrow"), and features (e.g., "tebestrian" for "pedestrian"). (Interestingly, she found no metathesis of consonants and vowels: phone-length metathesis across syllable boundaries always involved exchange of segments having similar functions within the syllable.) Fromkin concludes that an adequate model of speech performance must include mechanisms for producing such errors. We may say the same, mutatis mutandis, of an adequate model of speech perception.

---

[14] For experiments on and discussions of segmentation, see Harris (1953); Lisker (1957); Wang and Peterson (1958); Peterson et al. (1958); Lisker et al. (1962); Ladefoged (1967).

Each unit mentioned has been validated in perception. The feature is the least intuitively obvious segment and has received most experimental attention. For the syllable, we have already cited Liberman et al. (1952) and Öhman (1966) (p. 22). We may add, from among many, the series of experiments reported by Kozhevnikov and Chistovich (1965: Ch. VI) and an experiment of Huggins (1964), in which he alternated speech rapidly between ears and found that the rate most disruptive to speech perception was close to the syllable rate. For the phonetic segment, Pike (1943: Ch. VII) provided cogent arguments, observing, for example, that phonetic transcriptions of experts, and even of those with little training, generally agree on the total number of segments in an utterance. Fry, also, remarked that "the existence and widespread use of alphabetic writing are an indication that a phonemic system and segmentation into phonemic units are features which find a ready response in speakers and listeners" (Fry, 1964:60). We may add the experimental evidence of Kozhevnikov and Chistovich (1965:217 ff.), who found that mean reaction time for transcribing the consonant from a spoken CV syllable could be as much as 100 msec less than for transcribing the vowel from the same syllable. The same result was reported by Savin and Bever (1970). There is also evidence for the phoneme as an encoding unit in short-term memory (Conrad, 1964; Wickelgren, 1966b; Sperling and Speelman, 1970).

Savin and Bever (1970), as Warren (in press), made another interesting observation: subjects responded consistently faster to syllable targets than to phoneme targets.[15] They concluded that "phonemes are identified only after some larger linguistic sequence of which they are parts" (p. 300).[16] They explain that phonemes are primarily "neither perceptual nor articulatory," but rather "psychological entities of a non-sensory, non-motor kind...in short, phonemes are abstract" (p. 301). Without entering into discussion of the boundaries between sensation and perception, we may agree with their last remark since, as earlier observed, even phonetic perception is abstract. But if this distinction is to explain the longer latency for phoneme than for syllable identification, we must infer that syllables are not abstract. Certainly, as we argue below, syllables (unlike phonetic segments and features) may exist as articulatory, acoustic, and auditory units. But, insofar as they are phonetic units, they too are abstract. How an entity (whether concrete or abstract) of which the existence and form are determined by discrete components can be perceived without prior extraction of at least some of those components is hard to imagine. But it is not hard to imagine that the extraction of these

---

[15] Kolers (1968b) reports similar results for tachistoscopically exposed words and the letters that compose them.

[16] Ladefoged (1967:147) has developed a similar argument, drawing an unfortunate analogy with typing. He points out that skilled typists type by the word, not by the letter. Certainly, skilled typing may be governed by an hierarchical system of temporo-spatial coordination that includes integrating commands for sequences of letters as units [MacNeilage (1964) has an elegant discussion of these matters]. But it is evident that the typist does type letter by letter and that his behavior, not to mention the typewriter, would quickly jam, if he did not.

components is normally so rapid, automatic, and unconscious that their conscious recovery is slow. In fact, this may also be true of syllables in running speech: one would not be surprised to learn that recognition of syllables took longer than recognition of the words that they compose. Differences in recovery time for the several phonemes of Savin and Bever remain, of course, to be explained. In any event, their study has added evidence for the psychological and, in our view, the perceptual reality of the phonetic segment.

Finally, for the perceptual reality of features below the level of the phonetic segment, a large body of experimental evidence exists. First, virtually all studies of synthetic speech continua (many were cited in the preceding section) in which a "phoneme boundary" is observed may be regarded as studies of feature boundaries, since segments on either side of the boundary differ by a single articulatory feature. Second, perceptual confusions among consonants or vowels heard under difficult listening conditions (through noise, through filters, or under dichotic competition) group themselves systematically: the more feature values two segments have in common, the more likely they are to be confused. This has been shown for consonants[17] (Miller and Nicely, 1955; Singh, 1966, 1969; Studdert-Kennedy and Shankweiler, 1970) and for vowels (Miller, 1956; Singh and Woods, 1970). In several of these experiments, features were shown to be approximately additive (independent). A third line of evidence comes from scaling studies. Hanson (1967), for example, used multidimensional scaling techniques to place nine synthetic Swedish vowels in a psychological space which proved to have two dimensions corresponding to the tonal features (grave/acute; compact/diffuse) of Jakobson et al. (1963) and a third dimension corresponding to no defined feature [cf., the three dimensions of Pols, van der Kamp and Plomp (1969) in their study of Dutch vowels]. For scaling the six stop consonants of English, Greenberg and Jenkins (1964) used magnitude estimation and found, among other results, that sounds differing on one feature were judged to be closer than sounds differing on two. Peters (1963) found that manner, voicing, and place of articulation (in this order) were the main determinants of similarity judgments among consonants.[18] There is also evidence for encoding of both consonants and vowels in short-term memory according to phonological features (Wickelgren, 1965, 1966a, 1969; Sales et al., 1969, and four previous papers by these authors, cited therein). Finally, we note that several of these studies evaluated different sets of features and their definitional terms (articulatory, auditory). Klatt (1968) has presented a quantitative method for evaluating binary features from confusion matrices and for estimating their independence. For the present discussion, however, it is enough to know that some set of features functions in perception.

Let us now return to the theme by recalling that, despite their perceptual reality, neither phonetic segments nor their component features have acoustic

---

[17] For consonants, filtering typically tends to damage place of articulation; reverberation or echo effects tend to damage manner; noise damages both manner and place (Fant et al., 1966).

[18] Ladefoged (1969) was able to predict, with almost perfect accuracy, speaker judgments of articulatory similarities among thirty words in each of twenty Bantu languages by counting numbers of shared feature values on an ad hoc set of binary features.

reality. Our task is therefore to understand how these abstract (physically nonexistent) entities achieve psychological reality. The syllable has a different status, since we may define it not only linguistically but also in articulatory and acoustic terms. Whatever the difficulties of defining its boundaries acoustically [see Malmberg (1955) for one of the few attempts], its general function in production as a carrier of phonetic information is fairly clear. The syllable is the unit of consonant/vowel coarticulation: it arises from imposition of a precisely timed and coordinated pattern of articulatory gestures (the multiple physical manifestations of phonetic features) upon a pulse of air.[19] As the word itself indicates, the speaker collapses discrete muscular movements into a pattern of overlap that forms a larger unit.[20] For perception, the new unit has a double function. First, it reduces the number of auditory segments emitted per unit time below the number of phonetic segments and so brings segment repetition rate within the temporal resolving power of the ear. This function has been discussed elsewhere (Studdert-Kennedy and Liberman, 1963; Studdert-Kennedy and Cooper, 1966; Liberman et al., 1967b). Second, and we dwell on this here, its function is to contrast, and so to permit the listener to detect, segments of sound.

Fant and his colleagues (Fant and Lindblom, 1961; Fant, 1962) are among the few researchers to recognize the importance of sound contrast for phonetic perception. Dissatisfied by the abstract nature of distinctive features, and by the difficulty of specifying their acoustic correlates, Fant undertook to work upwards from signal to message rather than downwards from message to signal. He has developed a system for describing spectrograms in terms of sound segments with boundaries determined by switching events in the speech production mechanism. He decomposes sound segments into sound features specifying production and speech wave characteristics for each. He carefully reiterates that neither sound segments nor sound features are isomorphic with phonetic segments or features. For a recent account, see Fant (1968:235-241).

The importance of Fant's work is that, by systematic analysis of the acoustic signal, it provides an objective account of the factors with which the perceptual mechanism has to work. His system permits precise division of the spectrographic pattern in frequency and in time and invites exploration of the perceptual importance of its segments by filtering and gating techniques (e.g., Öhman 1961a, b; Fant, et al., 1966). Such studies may correct or corroborate work with synthetic speech, which, Fant (1964) believes, tends to overemphasize single cues at the expense of the entire auditory pattern.

Returning now to the syllable as vehicle of sound contrast, we may illustrate with part of a study by Bondarko (1969). Her intent was to examine "the means by which [phonemes] are contrasted within the syllable" (1969:2),

---

[19] We are aware that the work of Ladefoged and his colleagues (Ladefoged 1967: Ch. I) forces modification of Stetson's (1951) chest pulse theory of the syllable. But the disagreement is over the physiological control mechanism not over the function of the syllable in production and perception.

[20] The linguistic function of the syllable, as carrier of stress, is incidental to its coarticulatory origin and can be as well performed by an isolated vowel or, in some circumstances, consonant.

in other words, to search for the acoustic basis of distinctive feature oppositions. But we need not accept the theoretical framework to be interested by her study. Adopting an approach reminiscent of Fant's, she defined five types of contrast that may occur between sound segments in Russian CV syllables: contrasts in fundamental frequency, duration, formant structure, intensity, and "locus" (F2 transition). She then examined spectrographically twenty-five syllable types [five consonant classes--voiced and voiceless stops and fricatives, and sonants (laterals, nasals)--followed by five vowels]. She classified each syllable type according to how many of her contrast types it carried: the results ranged from all five for voiceless stops to one (or two) for laterals and nasals. The recorded syllables were gathered into random test orders and presented to twenty-five subjects for identification. Many details are omitted from the present account (in particular, the different values of stress used), but the general outcome was clear: probability of correct identification declined as intrasyllable contrast declined.

The portion of the syllable most likely to be missed with decline in contrast was the vowel. This is not unexpected, since we know vowel recognition to be heavily dependent on acoustic context, although most studies have given their attention to effects over signal stretches longer than the syllable (e.g., Ladefoged and Broadbent, 1957; Fry et al., 1962; Hiki et al., 1968). Among studies of effects within the syllable is one by Fujimura and Ochiai (1963), who compared identification of Japanese vowels spoken in syllabic context with identification of 50 msec segments gated out of the vowel centers: identifications shifted in the absence of surrounding formant movements. In a related study with synthetic vowels, Lindblom and Studdert-Kennedy (1967) showed that identification of a particular vowel pattern varied as a function of the rate and direction of its surrounding formant transitions.

An attempt not simply to demonstrate, but to watch the development of, contrast within the syllable is made by the shadowing studies of Kozhevnikov and Chistovich (1965). An experimenter reads over a microphone a list of VCV (or CV) patterns with V fixed and C a stop consonant that varies from trial to trial. A listener, in another room, repeats each utterance as rapidly as possible. Contact electrodes (lips, artificial palate) and throat microphones provide oscillographic records of the two speakers' utterances. If VCV patterns are being used, it is found that soon after the first speaker releases the consonant, the shadower's articulators constrict. The place of constriction may be more or less random, but, as the speaker continues, the shadower adjusts the point of constriction, if necessary, and completes his gesture. The latency of the shadower's release is 100-150 msec from the time of the speaker's release or from the time that the shadower's articulators assume the correct point of constriction. This delay is far too short for a normal choice reaction time. Kozhevnikov and Chistovich argue that shadowing shortcuts higher-level processes to reveal the normal, involuntary sequence of states in phonetic perception. Each state is said to be a function of both the preceding one and of some change in the external signal (p. 231). Since the first nonrandom state must be a function of the external signal, the entire sequence of phonetic states is a function of the external sequence. This sequence takes phonetic effect through acoustic contrast between its segments.

In short, experiments support a view of the syllable as carrier of contrast. But the contrast is not, as distinctive feature theory might have it, between linguistic features in their acoustic manifestations: we have seen that there are few, if any, acoustic segments isomorphic with linguistic features. We must, therefore, read the epigraph to this chapter in a sense different from that intended by its authors: "for the study of speech sounds ...their linguistic function is decisive" (Jakobson et al., 1963:11). One linguistic function of the syllable is to provide a rhythmic acoustic signal within the temporal resolving power of the ear and to facilitate, by its inherent acoustic contrasts, exercise of the listener's capacities for auditory discrimination. Without such a signal, linguistic communication would not be possible.

We conclude, then, that, while study of the acoustic signal may lend insight into its auditory function, it will not lead us appreciably closer to an understanding of how auditory patterns are related to phonetic matrix. For this, we must start from the known message and examine its manifestation in the signal.

THE PERCEPTUAL PHONETIC MATRIX

A sizeable body of knowledge exists concerning acoustic cues for phonetic segments and their features (see earlier citations, and Mattingly, this volume). The relations are not one-to-one and, from an acoustic point of view, seem arbitrary. Some features are signaled by more than one cue; some cues signal more than one feature. A brief explosion, for example, may indicate both voicing and place of articulation in final stops. In initial stops, voicing may be cued by explosion energy, degree of aspiration, and first formant intensity. Each cue may be emphasized in synthetic speech an used as the principal cue to voicing (Liberman et al., 1952, 1961b). In natural speech multiple cues, scattered within the syllable over time and frequency, combine, and according to synthesis experiments, their perceptual weights vary with phonetic context (Hoffman, 1958; Ainsworth, 1968). But we may make sense of their arbitrary nature and varying perceptial weights by applying the acoustic theory of speech production. The relations between source, vocal tract transfer function, and signal are well understood (Fant, 1960; Flanagan, 1965), and for any apparently arbitrary collocation of acoustic events, we may specify the conditions of production. The conditions of production themselves, however, remain unexplained until we can relate them to their underlying phonetic (articulatory) features.

For the feature of voicing, research is approaching this level of explanation. In 1960, Fant suggested that the main factor underlying the voiced/voiceless distinction for stop consonants in initial position was "the instant of time when the vocal cords close for the production of the following voiced sound" (Fant, 1960:225). Lisker and Abramson, by spectrographic analysis of stops in eleven languages and by perceptual experiments with synthetic speech in three, have explicated Fant's suggestion (Lisker and Abramson, 1964a, b, 1967, 1970; Abramson and Lisker, 1965, 1970; also Lisker et al., 1969). Their work suggests that the disparate acoustic features of explosion energy, aspiration, and first-formant intensity may all be derived from the single, underlying articulatory variable of voice onset time, that is, the relative timing of closure release and the onset of laryngeal vibration.[20]

---

[20] A fourth cue, rapid pitch changes at the onset of voicing, though probably trivial in natural speech, may be deliberately exaggerated in synthetic speech

They write:

> Laryngeal vibration provides the periodic or quasi-periodic carrier that we call voicing. Voicing yields harmonic excitation of a low frequency band during closure, and of the full formant pattern after release of the stop. Should the onset of voicing be delayed until some time after the release, however, there will be an interval between release and voicing onset when the relatively unimpeded air rushing through the glottis will provide the turbulent excitation of a voiceless carrier commonly called aspiration. This aspiration is accompanied by considerable attenuation of the first formant, an effect presumably to be ascribed to the presence of the tracheal tube below the open glottis. Finally, the intensity of the burst, that is, the transient shock excitation of the oral cavity upon release of the stop, may vary depending on the pressures developed behind the stop closure where such pressures will in turn be affected by the phasing of laryngeal closure. Thus it seems reasonable to us to suppose that all these acoustic features, despite their physical dissimilarities, can be ascribed ultimately to actions of the laryngeal mechanisms. (Abramson and Lisker, 1965)

We have quoted this account at length because it provides a model for the reduction of an apparently incoherent set of acoustic cues to an underlying articulatory variable or phonetic feature. How far this approach may be carried with other features remains to be seen. Also open for the future is the degree to which this and other approaches in experimental phonetics may force a modification of the phonetic features posited by phonological theory, if that theory is to be given a physical base. Ladefoged (1966) argued that the list of distinctive features was already overextended at fifteen, largely because the specifications disregarded physiological constraints on their combination.

Our argument then is that only through their articulatory origin can the temporally scattered and contextually variable acoustic (and auditory) patterns of speech be understood. The listener, who begins life as a babbler, develops, by repeated association of articulatory controls with their auditory consequences, a "knowledge" of his phonetic capacity and of the use to which it is put in his native language (Weir, 1962). By imitation of the voices around him and through adult acceptance of his imitations, he learns too the relation between his own acoustic output and that of others who have larger vocal tracts. Thus, he learns to "infer" from phonetically adventitious components of a speaker's signal (such as overall fundamental frequency or frequency position of the third formant) characteristics of the tract that produced it and the instructions required by his own tract for a "matching" signal. There are even grounds for suspecting that he may be born with some "knowledge" of phonetic capacity. Lisker and Abramson (1964a, b) found that nine languages, some with two, some with three stop

---

and will then serve as an effective cue (Haggard et al., 1970). The relatively small pitch changes of natural speech are probably also attributable to voice onset time.

categories, implemented only three values of voice onset time and suggested that physiological constraints may underlie this nonrandom distribution. Recently, Eimas and his colleagues (Eimas et al., 1971) tested discrimination of the labial voice onset time continuum in one-month-old infants by tracing adaptation and recovery of sucking responses. The infants showed significantly higher discrimination of a 20 msec difference in voice onset time that straddled two of Lisker and Abramson's phonetic categories than of the same difference within a category.

But whatever its source, "knowledge" of the relations between auditory patterns and articulatory features is available to every speaker/listener, and through this knowledge, we hypothesize, he is able to resolve an intricate pattern into its simple origin. In light of our earlier discussion, we must suppose the resolution to be automatic and, probably, beyond conscious recovery. Precisely how it may be accomplished we will not here speculate. But the general argument is not new. Similar positions have been adopted by Stevens and by Liberman and their colleagues (Halle and Stevens, 1962; Stevens and Halle, 1967; Stevens and House, 1970, Liberman et al., 1967b). Chistovich and her co-workers (Kozhevnikov and Chistovich, 1965; Chistovich et al., 1968) have applied the model experimentally. Their shadowing studies bear on perception only if we take stage 2 (phonetic) to be an automatic, running analysis, with its final output an assemblage of segments and features that may serve, in production, as instructions to the articulatory mechanism and, in perception, as input to the phonological component. In short, the perceptual matrix is identical with the generative. Its columns (phonetic segments) and rows (phonetic features), although determined by man's physiological capacities, are not themselves part of his auditory and articulatory systems. Rather, they are abstract linguistic entities uniting the complementary communicative functions of speaking and listening.

The model is broad, stripped of detail that might invite experimental test. Until our knowledge of perceptual processes and their physiological correlates has vastly increased, it is likely to remain so: less empirical than protreptic.

NEURAL SPECIALIZATION FOR SPEECH PERCEPTION

Cats can discriminate between speech sounds. Dewson (1964) trained five cats, by operant procedures, to discriminate between sustained [u] and [i], spoken by a man (F_o = 136 Hz), and to transfer their learning in one fifth the number of original trials to the same vowels spoken by a woman (F_o = 219 Hz). Warfield et al., (1966) trained ten cats to discriminate between the words "cat" and "bat," spoken by a woman, and demonstrated by a gating procedure that discrimination was based on the initial protion of each syllable. The cats did not transfer their learning to other words beginning with the same phonetic segments.

We may doubt that cats can perceive speech. This is not simply because they do not know a language, but also because they are not physiologically equipped to do so. For over a century, evidence has been accumulating that the human brain is asymmetrically organized for language functions. Recently, experiments have demonstrated that "dominance" of one or the other of the cerebral hemispheres (usually the left) extends to mechanisms for the perception of speech.

Kimura (1961a) discovered that listeners, presented with triads of competing digits in opposite ears (i.e., dichotically), were better able to recall those presented to the right ear than those presented to the left. She attributed the effect to functional prepotency of contralateral over ipsilateral auditory pathways and to cerebral dominance for language (Kimura, 1961b); her interpretation has since been supported by many studies. Shankweiler and Studdert-Kennedy (1966, 1967) showed that the right-ear advantage for speech did not depend on meaning, since it could be obtained if subjects were asked to identify contrasting consonants in pairs of dichotically presented, synthetic, CV nonsense syllables. There was no right-ear advantage if the competing stimuli were vowels. A later study with natural, CVC nonsense syllables (Studdert-Kennedy and Shankweiler, 1970) confirmed these results and showed, by error analysis, right-ear advantages for voicing and place of articulation in stop consonants, a result confirmed by Haggard (1970). Other work has demonstrated specialization of the nonlanguage hemisphere for recognition and discrimination of nonspeech auditory patterns (Miller, 1962; Kimura, 1964, 1967; Benton, 1965; Chaney and Webster, 1965; Shankweiler, 1966; Curry, 1967; Vignolo, 1969; Darwin, 1969).

Models of the neural mechanisms underlying these effects are still fluid, and no firm account will be offered here.[21] We will, however, assume that ear advantages reflect a degree of functional cerebral asymmetry. Cerebral asymmetry, or dominance, for speech perception requires that some portion of the perceptual function be performed more efficiently by the dominant hemisphere. One aim of current dichotic speech research is to define that portion by determining the acoustic and psychological conditions of the right-ear advantage.

There are three broad possibilities. The dominant hemisphere may be specialized for: 1) response alone, 2) phonetic analysis and response, or 3) auditory analysis, phonetic analysis, and response. The first possibility was tentatively ruled out by Studdert-Kennedy and Shankweiler (1970); subjects' error patterns indicated that place of articulation and voicing features in stop consonants were independently extracted by a single center in the dominant hemisphere. Darwin (1971) also concluded that specialization was not simply for response.

Darwin pushed the analysis further by showing that, in synthetic fricatives, the right-ear advantage for place of articulation only occurs if the feature is signaled by a formant transition, while for voicing, the advantage occurs only if the fricative is followed by a vowel. This suggests that the dominant hemisphere may be superior at some stage of auditory analysis, being better equipped, perhaps, for detection of certain acoustic features of speech, such as rapid formant movement or voice onset.

An alternative, though not incompatible, interpretation is that the dominant hemisphere is skilled at extracting phonetic information under "difficult" conditions, such as those provided by the complex auditory pattern of coarticulated consonant and vowel. This interpretation meshes with the lack of a reliable ear advantage for vowels and with the further fact that, if vowels are presented under conditions of general phonetic ambiguity, a right-ear advantage appears. Darwin (1971) demonstrated an advantage for dichotically competing, synthetic vowels, if the test included utterances apparently formed by different-sized vocal tracts: the listener was then in doubt as to which vocal tract would be presented to a particular ear on any trial. But if all vowels on the test sounded to have come from the same vocal tract, no right-ear advantage emerged.

---

[21] For relevant data and for discussion of possible mechanisms see Kimura, 1961b, 1967; Gazzaniga and Sperry, 1967; Milner et al., 1968; Sparks and Geschwind, 1968; Darwin, 1969, 1971; Halwes, 1969; Studdert-Kennedy and Shankweiler, 1970.

There are parallels here with earlier studies. Continuous perception of vowels gives way to categorical perception, if attention is forced to phonetic qualities. The characteristic speech mode of judging loudness, in terms of physiological effort required for phonation, appears if signals are presented in a context that demands phonetic processing. The characteristic speech skill of the dominant hemisphere is evinced in perception of rapid or otherwise "difficult" auditory/phonetic patterns. In other words, the evidence is consistent with the view that the language-dominant hemisphere is superior to the minor hemisphere in its capacity to accomplish the phonetic analysis of stage 2.

However, the peculiarity of speech perception cannot be solely in the use to which we put an acoustic signal. What engages the phonetic processors? Most nonspeech sounds cannot be heard as speech. But to Tennyson's farmer, the pony's hooves sang "Property, property, property," and we hear the chaffinch call "chewink," despite his lack of formant structure (Thorpe, 1961). A nonspeech sound with rapid acoustic variations that suggest the fundamental consonant/vowel alternations of speech may, if other conditions dispose, engage the phonetic processors. But the specifications are vague. Research has only begun (Darwin, 1969) to exploit the double dissociation of left and right hemispheres, speech and nonspeech signals, as the thin end of an experimental wedge for separating signals that are, or can be, heard as speech from those which cannot and for defining their characteristics. The acoustic border beyond which phonetic, and perhaps specialized auditory, processors are involuntarily engaged is still undefined.

Finally, the discovery that phonetic perception is neurologically linked to language processes emphasizes the unity of language and its medium of expression. Despite the contrary views of some phoneticians and linguists, language is not independent of its medium. Language, as we know it, could not exist without speech, nor speech without sound. Mattingly and Liberman (1970) explore formal analogies between phonetic and syntactic structure and suggest that study of speech perception may throw light on more general linguistic processes. At the same time, it may ground man's characteristic mode of communication in his physiology: to study speech is to enlarge our understanding of the biological foundations of language.

REFERENCES

Abercrombie, David. 1967. Elements of General Phonetics. Chicago: Aldine.
Abramson, Arthur S. 1961. Identification and discrimination of phonemic tones. JAcS. 33.842(A).
Abramson, Arthur S. and Leigh Lisker. 1965. Voice onset time in stop consonants: Acoustic analysis and synthesis. 5th Intern. Congr. Acoust., Leige.

Abramson, Arthur S. and Leigh Lisker. 1970. Discriminability along the voicing continuum: Cross-language tests. Proc. 6th Intern. Congr. Phon. Sci. Prague: Academia.

Ainsworth, W.A. 1968. Perception of stop consonants in synthetic CV syllables. L & S. 11.139-155.

Bastian, Jarvis, Peter D. Eimas, and Alvin M. Liberman. 1961. Identification and discrimination of a phonemic contrast induced by silent interval. JAcS. 33.842(A).

Benton, Arthur L. 1965. The problem of cerebral dominance. Canad. Psychologist 6.332-348.

Blumenthal, A. 1967. Prompted recall of sentences. J. verb. Learn. verb. Behav. 6.203-206.

Bolinger, Dwight A. 1958. A theory of pitch accent in English. Word 14.109-149.

Bondarko, L.V. 1969. The syllable structure of speech and distinctive features of phonemes. Phonetica 20.1-40.

Borst, John M. 1956. Use of spectrograms for speech analysis and synthesis. J. Audio Eng. Soc. 4.14-23.

Brown, Roger W. and D.C. Hildum. 1956. Expectancy and the perception of syllables. Language 32.411-419.

Capranica, R.R. 1965. The Evoked Vocal Response of the Bullfrog. Cambridge, Mass.: M.I.T. Press.

Chaney, R.B. and J.C. Webster. 1965. Information in certain multidimensional signals. U.S. Navy Electron. Lab. Rep. No. 1339, San Diego, Calif.

Chomsky, Noam. 1966. Topics in the theory of generative grammar. In Current Trends in Linguistics. Vol. III. Thomas A. Sebeok, Ed., The Hague: Mouton and Co.

Chomsky, Noam and Morris Halle. 1968. The Sound Pattern of English. New York: Harper and Row.

Chomsky, Noam and George A. Miller. 1963. Introduction to the formal analysis of natural languages. In Handbook of Mathematical Psychology, Vol. III. R.D. Luce, R.R. Bush and E. Galanter, Eds. New York: Wiley.

Chistovich, Ludmilla A., C. Gunnar, M. Fant, A. De Serpa-Leitão, and P. Tjernlund. 1966a. Mimicking of synthetic vowels. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR2.

Chistovich, Ludmilla A., C. Gunnar, M. Fant, and A. De Serpa-Leitão. 1966b. Mimicking and perception of synthetic vowels: Part II. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR3.

Chistovich, Ludmilla A., A. Golusina, V. Lublinskaja, T. Malinnikova, and M. Zukova. 1968. Psychological methods in speech perception research. Z. Phon. Sprachwiss. u. Komm. Fschg. 21.102-106.

Conrad, R. 1964. Acoustic confusions in immediate memory. Brit. J. Psychol. 55.75-84.

Cooper, Franklin S. 1950. Spectrum analysis. JAcS. 22.761-762.

Curry, Frederick K.W. 1967. A comparison of left-handed and right-handed subjects on verbal and non-verbal dichotic tasks. Cortex 3.343-352.

Darwin, Christopher J. 1969. Auditory perception and cerebral dominance. Unpublished Ph.D. thesis, University of Cambridge.

Darwin, Christopher J. 1971. Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23. To appear.

Day, Ruth S. 1968. Fusion in dichotic listening. Unpublished Ph.D. thesis, Stanford University.

Day, Ruth S. 1969. Temporal order judgments in speech. Paper presented at 9th Annual Meeting of Psychonomic Society, St. Louis.

Day, Ruth S. 1970. Temporal order perception of a reversible phoneme cluster. JAcS. 48.95(A).

Day, Ruth S. and James E. Cutting. 1970a. Levels of processing in speech perception. Paper presented at 10th Annual Meeting of Psychonomic Society, San Antonio.

Day, Ruth S. and James E. Cutting. 1970b. Perceptual competition between speech and nonspeech. Paper presented at 80th meeting of Acoustical Society of America, Houston.

Delattre, Pierre C., Alvin M. Liberman, and Franklin S. Cooper. 1955. Acoustic loci and transitional cues for consonants. JAcS. 27.769-773.

Delattre, Pierre C., Alvin M. Liberman, Franklin S. Cooper, and Louis J. Gerstman. 1952. An experimental study of the acoustic determinants of vowel color: Observations one- and two-formant vowels synthesized from spectrographic patterns. Word 8.195-210.

Dewson, J.H. 1964. Speech sound discrimination by cats. Science 144. 555-556.

Eimas, Peter D. 1963. The relation between identification and discrimination along speech and non-speech continua. L & S. 6.206-217.

Eimas, Peter D., Einar R. Siqueland, Peter Jusczyk, and James Vigorito. 1971. Speech perception in early infancy. Science 171. 303-306.

Fant, C. Gunnar M. 1960. Acoustic Theory of Speech Production. The Hague: Mouton.

Fant, C. Gunnar M. 1962. Descriptive analysis of the acoustic aspects of speech. Logos 5.3-17.

Fant, C. Gunnar M. 1964. Auditory patterns of speech. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR3. 16-20.

Fant, C. Gunnar M. 1966. A note on vocal tract size factors and non-uniform F-pattern scalings. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR4.

Fant, C. Gunnar M. 1968. Analysis and synthesis of speech processes. In Manual of Phonetics. Bertil Malmberg, Ed. Amsterdam: North-Holland Publishing Company, 173-277.

Fant, C. Gunnar M. and Bjorn E.F. Lindblom. 1961. Studies of minimal speech sound units. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR2.

Fant, C. Gunnar M., Bjorn E.F. Lindblom, and A. De Serpa-Leitão. 1966. Consonant confusions in English and Swedish. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR4.

Flanagan, James L. 1965. Speech Analysis, Synthesis and Perception. New York: Academic Press.

Fodor, Jerry A. and Thomas G. Bever. 1965. The psychological reality of linguistic segments. J. verb. Learn. verb. Behav. 4.414-420.

Fourcin, Adrian J. 1968. Speech source inference. IEEE Transactions on Audio- and Electroacoustics. AV-16.65-67.

Frishkopf, Larry S. and Moise H. Goldstein, Jr. 1963. Responses to acoustic stimuli from single units in the eighth nerve of the bullfrog. JAcS. 35. 1219-1228.

Fromkin, Victoria A. 1966. Neuromuscular specification of linguistic units. L & S. 9.170-199.

Fromkin, Victoria A. 1970. Tips of the slung or to err is human. Univ. of California at Los Angeles: Working Papers in Phonetics. 14.40-79.

Fry, Dennis B. 1955. Duration and intensity as physical correlates of linguistic stress. JAcS. 27.765-768.

Fry, Dennis B. 1956. Perception and recognition in speech. In For Roman Jakobson. Morris Halle, Horace G. Lunt, Hugh McClean, and Cornelis H. van Schooneveld, Eds. 169-173. The Hague: Mouton.

Fry, Dennis B. 1964. Experimental evidence for the phoneme. In In Honor of Daniel Jones. David Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott, and J.L.M. Trim, Eds. 59-72. London: Logmans.

Fry, Dennis B. 1968. Prosodic phenomena. In Manual of Phonetics. Bertil Malmberg, Ed. 365-411. Amsterdam: North-Holland Publishing Company.

Fry, Dennis B., Arthur S. Abramson, Peter D. Eimas, and Alvin M. Liberman. 1962. The identification and discrimination of synthetic vowels. L & S. 5.171-189.

Fujimura, O. and K. Ochiai. 1963. Vowel identification and phonetic contexts. JAcS. 35.1889(A).

Fujisaki, Hiroya and Takako Kawashima. 1968. The roles of pitch and higher formants in the perception of vowels. IEEE Transactions on Audio- and Electro-acoustics. AV-16.73-77.

Fujisaki, Hiroya and Takako Kawashima. 1969. On the modes and mechanisms of speech perception. Division of Electrical Engineering, Engineering Research Institute, University of Tokyo, Annual Report No. 1.67-73.

Fujisaki, Hiroya and Naoshi Nakamura. 1969. Normalization and recognition of vowels. Division of Electrical Engineering, Engineering Research Institute, University of Tokyo, Annual Report No. 1.61-66.

Garrett, Merrill, Thomas G. Bever, and Jerry A. Fodor. 1966. The active use of grammar in speech perception. Percept. and Psychophysics 1.30-32.

Gazzaniga, Michael S. and Roger W. Sperry. 1967. Language after section of the cerebral commissures. Brain 90.131-148.

Gerstman, Louis J. 1968. Classification of self-normalized vowels. IEEE Transactions on Audio- and Electro-acoustics. AV-16.78-80.

Greenberg, Joshuah J. and James J. Jenkins. 1964. Studies in the psychological correlates of the sound system of American English. Word 20.157-177.

Hadding-Koch, Kerstin. 1961. Acousticophonetic studies in the intonation of Southern Swedish. Travaux de L'Institut de Phonetique de Lund III. Lund: C.W.K. Gleerup.

Hadding-Koch, Kerstin and Michael Studdert-Kennedy. 1963. A study of semantic and psychophysical test responses to controlled variations in fundamental frequency. Studia Linguistica 17.65-76.

Hadding-Koch, Kerstin and Michael Studdert-Kennedy. 1964. An experimental study of some intonation contours. Phonetica 11.175-185.

Haggard, Mark P. 1970. The use of voicing information. Psychological Laboratory, University of Cambridge, Progress Report No. 2.1-14.

Haggard, Mark P., Stephen Ambler, and Mo Callow. 1970. Pitch as a voicing cue. JAcS. 47.613-617.

Halle, Morris. 1964. On the bases of phonology. In The Structure of Language, Jerry A. Fodor, and Jerrold J. Katz, Eds. Englewood Cliffs, New Jersey: Prentice-Hall.

Halle, Morris and Kenneth N. Stevens. 1962. Speech recognition: A model and a program for research. IRE Transactions on Information Theory IT-8. 155-159.

Halwes, Terry G. 1969. Effects of dichotic fusion in the perception of speech. Unpublished Ph.D. thesis, University of Minnesota.

Hanson, G. 1967. Dimensions in speech sound perception: An experimental study of vowel perception. Ericsson Tech. 23.3-173.

Harris, Cyril M. 1953. A study of the building blocks of speech. JAcS. 25. 962-969.

Harris, Katherine S. Physiological aspects of articulatory behavior. See this volume.

Harris, Katherine S., Malcolm H. Schvey, and Gloria F. Lysaught. 1962. Component gestures in the production of oral and nasal labial stops. JAcS. 34.743(A).

Harris, Katherine S., Gloria F. Lysaught, and Malcolm H. Schvey. 1965. Some aspects of the production of oral and nasal labial stops. L & S. 8. 135-147.

Hiki, Shizuo, Hiroshi Sato, Takeo Igarashi, and Juro Oizumi. 1968. Dynamic model of vowel perception. Paper presented at Sixth Internat. Congr. on Acoustics, Tokyo.

Hockett, Charles F. 1958. A Course in Modern Linguistics. New York: Macmillan.

Hoffman, Howard S. 1958. Study of some cues in the perception of voiced stop consonants. JAcS. 30.1035-1041.

Huggins, A.W.F. 1964. Distortion of the temporal pattern of speech: Interruption and alternation. JAcS. 36.1055-1064.

Jakobson, Roman, Gunnar Fant, and Morris Halle. 1963. Preliminaries to speech analysis. Cambridge, Mass.: M.I.T. Press.

Jakobson, Roman and Morris Halle. 1956. Fundamentals of Language. The Hague: Mouton.

Johnson, N. 1966. The psychological reality of phrase structure rules. J. verb. Learn. verb. Behav. 4.469-475.

Joos, Martin A. 1948. Acoustic phonetics. Language, Suppl. 24.1-136.

Kasuya, H., H. Suzuki, and K. Kido. 1967. Acoustic parameters necessary to discriminate the vowels. Reports of Research Institute of Electrical Communication. Tōhoku University. Vol. 19, No. 3.

Kimura, Doreen. 1961a. Some effects of temporal lobe damage on auditory perception. Canad. J. Psychol. 15.156-165.

Kimura, Doreen. 1961b. Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15.166-171.

Kimura, Doreen. 1964. Left-right differences in the perception of melodies. Quart. J. Exp. Psychol. 16.355-358.

Kimura, Doreen. 1967. Functional asymmetry of the brain in dichotic listening. Cortex 3.163-178.

Kirstein, Emily. 1970. Unpublished Ph.D. thesis, University of Connecticut, Storrs.

Klatt, Dennis H. 1968. Structure of confusions in short-term memory between English consonants. JAcS. 44.401-407.

Klatt, Dennis H. 1969. Perception of linguistic and phonetic units. Paper presented at the session on Speech Synthesis and Speech Perception, Annual Meeting of the American Association for the Advancement of Science, December 28th, Boston, Massachusetts.

Koenig, W.H., H.K. Dunn, and L.Y. Lacey. 1946. The sound spectrograph. JAcS. 18.19-49.

Kolers, Paul A. 1968a. Some psychological aspects of pattern recognition. In Recognizing Patterns. Paul A. Kolers and M. Eden, Eds. Cambridge, Mass.: M.I.T. Press.

Kolers, Paul A. 1968b. Reading temporally and spatially transformed text. In The Psycholinguistic Nature of the Reading Process. Kenneth S. Goodman, Ed. Detroit: Wayne University Press.

Kozhevnikov, V.A. and Ludmilla A. Chistovich. 1965. Rech' Artikuliatsia i vospriiatie. Moscow-Leningrad. Trans. as Speech, Articulation and Perception. Washington: Clearinghouse for Federal Scientific and Technical Information. JPRS 30-543.

Ladefoged, Peter. 1966. The nature of general phonetic theories. In Languages and Linguistics. Georgetown University: Monograph No. 18, 27-42.

Ladefoged, Peter. 1967. Three Areas of Experimental Phonetics. New York: Oxford University Press.

Ladefoged, Peter. 1969. The measurement of phonetic similarity. Paper presented at International Conference on Computational Linguistics, Stockholm, Sweden.

Ladefoged, Peter and Donald E. Broadbent. 1957. Information conveyed by vowels. JAcS. 29.98-104.

Lane, Harlan L. 1965. The motor theory of speech perception: A critical review. Psych. Rev. 72.275-309.

Lane, Harlan L., A.C. Catania, and S.S. Stevens. 1961. Voice level: Autophonic scale, perceived loudness, and effects of sidetone. JAcS. 33. 160-167.

Lehiste, Ilse. 1970. Suprasegmentals. Cambridge, Mass.: M.I.T. Press.

Lehiste, Ilse and Gordon E. Peterson. 1959. Vowel amplitude and phonemic stress in American English. JAcS. 31.428-435.

Liberman, Alvin M. 1957. Some results of research on speech perception. JAcS. 29.117-123.

Liberman, Alvin M., Franklin S. Cooper, Katherine S. Harris, Peter F. MacNeilage, and Michael Studdert-Kennedy. 1967a. Some observations on a model for speech perception. In Models for the Perception of Speech and Visual Form. Weiant Wathen-Dunn, Ed. Cambridge, Mass.: M.I.T. Press.

Liberman, Alvin M., Franklin S. Cooper, Donald Shankweiler and Michael Studdert-Kennedy. 1967b. Perception of the speech code. Psychol. Rev. 74.431-461.

Liberman, Alvin M., Pierre C. Delattre, and Franklin S. Cooper. 1952. The role of selected stimulus variables in the perception of the unvoiced stop consonants. Amer. J. Psychol. 65.497-516.

Liberman, Alvin M., Pierre C. Delattre, Franklin S. Cooper, and Louis J. Gerstman. 1954. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monograph. 68 (whole number).

Liberman, Alvin M., Katherine S. Harris, Peter D. Eimas, Leigh Lisker, and Jarvis Bastian. 1961a. An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. L & S. 4.175-195.

Liberman, Alvin M., Katherine S. Harris, Howard S. Hoffman, and Belver C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54.358-368.

Liberman, Alvin M., Katherine S. Harris, Joanne Kinney, and Harlan Lane. 1961b. The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. 61.379-388.

Licklider, Joseph C.R. and George A. Miller. 1951. The perception of speech. In Handbook of Experimental Psychology. S.S. Stevens, Ed. New York: Wiley.

Lieberman, Philip. 1963. Some effects of semantic and grammatical context on the production and perception of speech. L & S. 6.172-179.

Lieberman, Philip. 1965. On the acoustic basis of the perception of intonation by linguists. Word 21.40-54.

Lieberman, Philip. 1967. Intonation, Perception and Language. Cambridge, Mass.: M.I.T. Press.

Lieberman, Philip. 1970. Toward a unified phonetic theory. Linguistic Inquiry 1.307-322.

Lindblom, Bjorn. 1963. Spectrographic study of vowel reduction. JAcS. 35. 1773-1781.

Lindblom, Bjorn E.F. and Michael Studdert-Kennedy. 1967. On the role of formant transitions in vowel recognition. JAcS. 42.830-843.

Lisker, Leigh. 1957. Linguistic segments, acoustic segments and synthetic speech. Language 33.370-374.

Lisker, Leigh and Arthur S. Abramson. 1964a. A cross-language study of voicing in initial stops: Acoustical measurements. Word 20.384-422.

Lisker, Leigh and Arthur S. Abramson. 1964b. Stop categorization and voice onset time. Proc. 5th Internat. Congr. of Phonetic Sciences, Munster.

Lisker, Leigh and Arthur S. Abramson. 1967. Some effects of context on voice on set time in English stops. L & S. 10.1-28.

Lisker, Leigh and Arthur S. Abramson. 1970. The voicing dimension: Some experiments in comparative phonetics. Proc. 6th Intern. Congr. Phon. Sci. Prague: Academia.

Lisker, Leigh, Arthur S. Abramson, Franklin S. Cooper, and Malcolm H. Schvey. 1969. Transillumination of the larynx in running speech. JAcS. 45. 1544-1546.

Lisker, Leigh, Franklin S. Cooper, and Alvin M. Liberman. 1962. The uses of experiment in language description. Word 18.82-106.

Lotz, John, Arthur S. Abramson, Louis J. Gerstman, Frances Ingemann, and William J. Nemser. 1960. The perception of English stops by speakers of English, Spanish, Hungarian and Thai: A tape-cutting experiment. L & S. 3.71-77.

MacNeilage, Peter F. 1963. Electromyographic and acoustic study of the production of certain final clusters. JAcS. 35.461-463.

MacNeilage, Peter F. 1964. Typing errors as clues to serial order mechanisms in language behavior. L & S. 7.144-159.

MacNeilage, Peter F. 1970. Motor control of serial ordering of speech. Psych. Rev. 77.182-196.

MacNeilage, Peter F. and Joseph L. De Clerk. 1969. On the motor control of coarticulation in CVC monosyllables. JAcS. 45.1217-1233.

Malécot, Andre. 1956. Acoustic cues for nasal consonants. Language 32. 274-284.

Malmberg, Bertil. 1955. The phonetic basis for syllable division. Studia Linguistica 9.80-87.

Mattingly, Ignatius G. 1966. Synthesis by rule of general American English. Supplement to Haskins Laboratories Status Report.

Mattingly, Ignatius G. Speech synthesis for phonetic and phonological models. See this volume.

Mattingly, Ignatius G. and Alvin M. Liberman. 1970. The speech code and the physiology of language. In Information Processing in the Nervous System. K.N. Leibovic, Ed. New York: Springer. 97-117.

Mattingly, Ignatius G., Alvin M. Liberman, Ann K. Syrdal, and Terry G. Halwes. In press. Discrimination in speech and nonspeech modes. Perception and Psychophysics.

Mehler, J. 1963. Some effects of grammatical transformations on the recall of English sentences. J. verb. Learn. verb. Behav. 2.346-351.

Menon, K.M.N., Paul J. Jensen, and Donald Dew. 1969. Acoustic properties of certain VCC utterances. JAcS. 46.449-457.

Miller, George A. 1951. Language and Communication. New York: McGraw Hill.

Miller, George A. 1956. The perception of speech. In For Roman Jakobson. Morris Halle, Horace G. Lunt, Hugh McClean and Cornelis H. van Schooneveld, Eds. The Hague: Mouton. 353-359.

Miller, George A. 1964. Some psychological studies of grammar. Amer. Psychol. 17.748-762.

Miller, George A., G.A. Heise, and W. Lichten. 1951. The intelligibility of speech as a function of the context of the test materials. JAcS. 41.329-335.

Miller, George A. and S. Isard. 1963. Some perceptual consequences of linguistic rules. J. verb. Learn. verb. Behav. 2.217-228.

Miller, George A. and Patricia Nicely. 1955. An analysis of some perceptual confusions among some English consonants. JAcS. 27.338-352.

Milner, Brenda. 1962. Laterality effects in audition. In Interhemispheric Relations and Cerebral Dominance. V.B. Montcastle, Ed. Baltimore: Johns Hopkins Univ. Press.

Milner, Brenda, L. Taylor, and Roger W. Sperry. 1968. Lateralized suppression of dichotically-presented digits after commissural section in man. Science 161.184-185.

Neisser, Ulric. 1966. Cognitive Psychology. New York: Appleton-Century-Crofts.

Öhman, Sven E.G. 1961a. On the contribution of speech segments to the identification of Swedish consonant phonemes. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR2.

Öhman, Sven E.G. 1961b. Relative importance of sound segments for the identification of Swedish stops in VC and CV syllables. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR3.

Öhman, Sven E.G. 1965. On the coordination of articulatory and phonatory activity in the production of Swedish tonal accents. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR2.

Öhman, Sven E.G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. JAcS. 39.151-168.

Öhman, Sven E.G. 1967. Numerical model of coarticulation. JAcS. 41.310-320.

Öhman, Sven E.G., A. Persson, and R. Leanderson. 1967. Speech production at the neuromuscular level. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR2-3.

Peters, R.W. 1963. Dimensions of perception for consonants. JAcS. 35.1985-1989.

Peterson, Gordon E. 1952. Information bearing elements of speech. JAcS. 24.629-637.

Peterson, Gordon E. 1959. Vowel formant measurements. J. Speech and Hearing Res. 2.173-183.

Peterson, Gordon E. 1961. Parameters of vowel quality. J. Speech and Hearing Res. 4.10-29.

Peterson, Gordon E. and H.L. Barney. 1952. Control methods used in a study of vowels. JAcS. 25.175-184.

Peterson, Gordon E., William S-Y. Wang, and E. Sivertsen. 1958. Segmentation techniques of speech synthesis. JAcS. 30.739-742.

Pickett, James M. and Irwin Pollack. 1963. Adjacent context and the intelligibility of words excised from fluent speech. JAcS. 35.807(A).

Pike, Kenneth L. 1943. Phonetics. Ann Arbor: Univ. of Michigan Press.

Pols, L.C.W., L.J. Th. van der Kamp, and R. Plomp. 1969. Perceptual and physical space of vowel sounds. JAcS. 46.458-467.

Potter, Ralph K., George A. Kopp, and Harriet C. Green. 1947. Visible Speech. New York: Van Nostrand.

Reber, A.S. and J.R. Anderson. 1970. The perception of clicks in linguistic and non-linguistic messages. Percept. and Psychophysics. 8.81-89.

Sachs, Richard M. 1969. Vowel identification and discrimination in isolation vs. word context. Research Laboratory of Electronics, M.I.T., QPR93. 220-229.

Sales, Bruce D., Ronald A. Cole, and Ralph N. Haber. 1969. Mechanisms of aural encoding: V. Environmental effects of consonants on vowel encoding. Percept. and Psychophysics. 6.361-365.

Savin, Harris B. 1963. Word-frequency effect and errors in the perception of speech. JAcS. 35.200-206.

Savin, Harris B. and Thomas G. Bever. 1970. The non-perceptual reality of the phoneme. J. verb. Learn. verb. Behav. 9.295-302.

Savin, Harris B. and E. Perchonock. 1965. Grammatical structure and the immediate recall of English sentences. J. verb. Learn. verb. Behav. 4. 348-353.

Schatz, Carol. 1954. The role of context in the perception of stops. Language 30.47-56.

Shankweiler, Donald. 1966. Effects of temporal-lobe damage on perception of dichotically presented melodies. J. Comp. Physiol. Psychol. 62.115-119.

Shankweiler, Donald and Michael Studdert-Kennedy. 1966. Lateral differences in perception of dichotically presented synthetic consonant-vowel syllables and steady-state vowels. JAcS. 39.1256(A).

Shankweiler, Donald and Michael Studdert-Kennedy. 1967. Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19.59-63.

Shearme, J.N. and J.N. Holmes. 1962. An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1-formant 2 plane. In Proc. of Fourth Internat. Congr. of Phonet. Sci., Helsinki, 1961. The Hague: Mouton. 234-240.

Singh, Sadanand. 1966. Cross-language study of perceptual confusions of plosive phonemes in two conditions of distortion. JAcS. 40.635-656.

Singh, Sadanand. 1969. Interrelationship of English consonants. In Proc. 6th Internat. Congr. Phonetic Sci. Prague: Academia. 542-544.

Singh, Sadanand and D. Woods. 1970. Multidimensional Scaling of 12 American English vowels. JAcS. 48.104(A).

Skinner, Burrhus F. 1936. The verbal summator and a method for the study of latent speech. J. Psychol. 2.71-107.

Sparks, Robert and Norman Geschwind. 1968. Dichotic listening in man after section of neocortical commissures. Cortex 4.3-16.

Sperling, George and Roseanne G. Speelman. 1970. Acoustic similarity and short-term memory: Experiments and a model. In Models of Human Memory. Donald A. Norman. New York: Academic Press. 151-202.

Stetson, R.H. 1951. Motor Phonetics. Amsterdam: North-Holland.

Stevens, Kenneth N. 1967. Acoustic correlates of certain consonantal features. Paper presented at Conference on Speech Communication and Processing, Cambridge, Mass.

Stevens, Kenneth N. 1968a. Acoustic correlates of place of articulation for stop and fricative consonants. QPR. 89. Research Laboratory of Electronics, M.I.T., 199-205.

Stevens, Kenneth N. 1968b. On the relations between speech movements and spech perception. Z. Phon., Sprachwiss. u. Komm. Fschg. 21.102-106.

Stevens, Kenneth N. In press. The quantal nature of speech: Evidence from articulatory-acoustic data. In Human Communication: A Unified View. E.E. David and P.B. Denes, Eds.

Stevens, Kenneth N. and Morris Halle. 1967. Remarks on analysis by synthesis and distinctive features. In Models for the Perception of Speech and Visual Form. Weiant Wathen-Dunn, Ed. Cambridge, Mass.: M.I.T. Press.

Stevens, Kenneth N. and Arthur S. House. 1963. Perturbation of vowel articulations by consonantal context. J. Speech and Hearing Research 6.111-128.

Stevens, Kenneth N. and Arthur S. House. 1970. Speech perception. In Foundations of Modern Audiometry. J. Tobias and E. Schubert, Eds.

Stevens, Kenneth N., Arthur S. House, and A.P. Paul. 1966. Acoustical description of syllabic nuclei. JAcS. 40.123-132.

Stevens, Kenneth N., Alvin M. Liberman, Michael Studdert-Kennedy, and Sven E.G. Ohman. 1969. Cross-language study of vowel perception. L & S. 12.1-23.

Studdert-Kennedy, Michael and Franklin S. Cooper. 1966. High-performance reading machines for the blind: Psychological problems, technological problems and status. Proc. Internat. Confer. on Sensory Devices for the Blind, St. Dunstan's, London. 317-340.

Studdert-Kennedy, Michael and Kerstin Hadding. In preparation. Further experimental studies of fundamental frequency contours.

Studdert-Kennedy, Michael and Alvin M. Liberman. 1963. Psychological considerations in the design of auditory displays for reading machines. Proc. Internat. Congr. on Technology and Blindness. 289-304.

Studdert-Kennedy, Michael and Donald Shankweiler. 1970. Hemispheric specialization for speech perception. JAcS. 48.579-594.

Studdert-Kennedy, Michael, Donald Shankweiler, and Susan Schulman. 1970a. Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. JAcS. 48.599-602.

Studdert-Kennedy, Michael, Alvin M. Liberman, Katherine S. Harris, and Franklin S. Cooper. 1970b. The motor theory of speech perception: A reply to Lane's critical review. Psych. Rev. 77.234-249.

Tatham, Marcel A.A. and Katherine Morton. 1968. Some electromyography data towards a model of speech production. Univ. of Essex Language Center, Occasional Papers 1.1-24.

Thorpe, W.H. 1961. Bird-Song. Cambridge, England: Cambridge University Press.

Vignolo, Luigi A. 1969. Auditory agnosia: A review and report of recent evidence. In Contributions to Clinical Neuropsychology. Arthur Benton, Ed. Chicago: Aldine. 172-208.

Wang, William S-Y and Charles Fillmore. 1961. Intrinsic cues and consonant perception. J. Speech and Hearing Research 4.130-136.

Wang, William S-Y and Gordon E. Peterson. 1958. Segment inventory for speech synthesis. JAcS. 30.743-746.

Warfield, D., R.J. Ruben, and R. Glackin. 1966. Word discrimination in cats. J. Audit. Res. 6.97-119.

Warren, Richard M. 1970. Perceptual restoration of missing speech sounds. Science 167.392-393.

Warren, Richard M. In press. Identification times for phonemic components of graded complexity and for spelling of speech. Perception and Psychophysics.

Webster, John C. 1969. Effects of noise on speech intelligibility. In Noise as a Public Health Hazard. W.D. Ward and J.E. Fricke, Eds. Washington: American Speech and Hearing Association Reports 4.

Weir, Ruth. 1962. Language in the Crib. The Hague: Mouton.

Whitfield, I.C. and E.F. Evans. 1965. Responses of auditory cortical neurons to stimuli of changing frequency. J. Neurophysiology 28.655-672.

Wickelgren, Wayne A. 1965. Distinctive features and errors in short-term memory for English vowels. JAcS. 38.583-588.

Wickelgren, Wayne A. 1966a. Distinctive features and errors in short-term memory for English consonants. JAcS. 39.388-398.

Wickelgren, Wayne A. 1966b. Short-term recognition memory for single letters and phonemic similarity of retroactive interference. Quart. J. Exptl. Psychol. 18.55-62.

Wickelgren, Wayne A. 1969. Auditory or articulatory coding in verbal short-term memory. Psychol. Rev. 76.232-235.

Physiological Aspects of Articulatory Behavior*

Katherine S. Harris[+]
Haskins Laboratories, New Haven

INTRODUCTION

The motor theory of speech perception is a statement that we will find a simpler relationship between the string of phonemes that a listener perceives and the articulation of the speaker than between the acoustic signal the speaker generates and perception.

The inconsistencies in the acoustic signal from speaker to speaker and within the speech of a given speaker have three different types of causes--differences in vocal tract size and shape among different speakers, apparent differences in style of different speakers of the same language, and differences in the production of a given phoneme by the same speaker in different contexts.

In principle, the acoustic differences produced by differences in vocal-tract size are quite well understood through a long tradition of study which led to Fant's Acoustic Theory of Speech Production (1960). The modern acoustic theory of speech production permits us to calculate the output sound when the cross-sectional area of the vocal tract is known, and vice versa. Apparently, listeners make some such calculation in extracting messages from speech, and this must be, in part, what enables little children to imitate the speech of adults, even though the shape changes which occur during growth are quite complicated.

Stylistic differences between speakers are of several sorts, having different origins and consequences. First, there are dialectal differences in such things as vowel systems, and second, there is the large class of idiosyncracies that are lumped as speech defects; we will not be concerned with these two sources of phoneme difference further in this paper, however troublesome it may be in practice to sort them out from those discussed.

Our chief concern here will be to discuss differences between contexts in the production of a given phoneme by a single speaker. Can these allophonic differences be assigned to a single phoneme which is present at some level in the nervous system and is variably executed at lower articulatory levels? Furthermore, are phonemes the smallest units of speech storage, or can they be considered, in turn, to be combinations of still smaller invariant parts?

---

*Chapter prepared for Current Trends in Linguistics, Vol. XII, Thomas A. Sebeok Ed. (The Hague: Mouton).

[+] Also, City University of New York Graduate Center, New York.

This rather general question can be rephrased as a more specific hypothesis. It has sometimes been assumed that each phoneme is stored in the nervous system as a fixed articulatory position, or target. This target is not always reached, but failures of attainment are a consequence of the way the articulatory apparatus operates. Let us explore the operation, to test the truth of the assumption itself.

There have been three general modes of attack on the problem, when phrased in this way. The first is purely acoustic. Both Lindblom (1963) and Öhman (1966) have used acoustic data to make inferences about the nature of shape differences between allophones. This procedure works because the acoustic theory of speech production will generally allow direct inference from acoustic output to shapes, at least within a single speaker. The only problem is that the relationship between acoustic output and articulator shape is sharply nonlinear--that is, very small changes in shape from some positions will produce large changes in acoustic output, while in other positions, large changes in shape will produce only small changes in acoustic output. Indeed, this property of the shape-to-acoustic transform has led Stevens (1969) to propose that the articulatory positions of the phonemes have been chosen at locations where small changes in articulation will cause minimal changes in acoustic effect.

A second way of investigating the problem is to look at the movements of the articulatory organs directly. Some methods, such as the observation of the up-and-down movements of the larynx, have a long history in experimental phonetics. Some, such as cinefluorography, are recently developed. All of these are alike in that they are observations of the last stage in the articulatory process before transformation of the articulatory signal into sound and, hence, are equivalent to the acoustic methods discussed above.

A third general technique, also physiological, is examination of the myographic signals generated by the muscles as they contract. This technique is different from those described above in that the signal observed is related not to the position assumed by the articulator but to the force acting on the articulator to bring it to a certain position. A brief description of the electromyographic signal may perhaps make this point clearer.

The muscles of the body are made up of bundles of fibers, organized into what are called motor units. The fibers in each motor unit contract when they receive an impulse from the single nerve fiber which supplies them. Both the nerve impulse and the muscle contraction are accompanied by electrical potential, but the muscle potential, a relatively large signal, is what is observed in ordinary electromyographic (EMG) recording. The essentials of a recording device are an electrode sensitive to differences in electrical potential between two points, an amplifying device, and some form of recorder, which shows a transformation of the potential difference as a function of time.

The electromyographic record shown in the typical study described below is a record of the output of a large number of motor units. In general, the fluctuations of amplitude as a function of time are related to changes in the strength of muscle contraction, but their relationship is by no means a simple one, partly because a stronger contraction is accompanied both by change in the number of motor units firing and by changes in the frequency with which they fire. Furthermore, the size of the recording electrode and its distance from the active muscle fibers will affect the record obtained.

For a somewhat more detailed discussion of the recording techniques used in speech research, the curious reader can see Harris (in press) or, for a discussion of general electromyographic technique, a general text such as Basmajian (1962). It is customary nowadays to present records after computer processing. However, in spite of the trying problems in arranging for such a result (again, see Harris, in press, or Cooper, 1965), the function of the computer is just to provide simple averaging of repeated utterances.

Returning to the relationship between articulator shape and electromyographic signal, it is clear that the transformation that takes place between them is quite complex. As an example, let us assume a given vowel, [i], is represented by a fixed articulatory position (an assumption which is only approximately true). Since electromyographic signals are related to the force of muscle contraction, a movement to the vowel from a consonant with a similar place of articulation, as in the sequence [ti], will involve a smaller signal than from an open vowel, as in the sequence [ai]. This point is made very clearly by MacNeilage (1970) and will be discussed further below. Another complication arising from the nature of EMG signals is that variation in size will be associated both with speed of articulator movement and with distance. Generally, a larger signal is associated with a faster, as well as a larger, movement.

In spite of these complications, there seem to be three rather compelling reasons for studying the motor patterns of articulation. First, as Cooper (1965) and Liberman et al. (1967) have pointed out, they are one step closer to the speech-generating center in the brain than are articulatory shapes. Second, most articulators are rather inaccessible; in some cases, it may be easier to examine EMG signals than other physiological variables. Third, even when only anatomical data about the muscles are available, they allow some insight into the articulatory shapes that the speaker can generate.

This diversity of uses of muscle study has resulted in a diversity of purpose for studies already performed; some are aimed at simply describing the mechanism for a given action and some at the more general purpose of understanding speech generation. In what follows, we will try to sketch the anatomy of the articulatory muscles and then to discuss what EMG studies, and the other closely related physiological research, have to offer for understanding the organization of speech. For anatomical detail, the interested reader can refer to more general speech textbooks, such as Zemlin (1968) or van Riper and Irwin (1958) or standard anatomical works.

THE ORGANIZATION OF THE SPEECH MUSCULATURE

The speech musculature can be divided into three more or less independent groups--respiratory, laryngeal, and articulatory.

The respiratory muscles act to provide power for the speech mechanism, the laryngeal muscles act to transform the power into acoustic energy, and the articulatory muscles modulate this carrier to produce the specific sounds of speech, though the functions of these three groupings overlap. We will discuss each of the three in turn.

The Respiratory System

In normal breathing, inspiration and expiration occupy approximately equal amounts of time. In speech, the inspiration-expiration ratio is considerably changed, so that the expiration occupies about seven-eights of the total cycle. The mechanisms of breathing seem to set about four seconds as the limit on an expiration, without a necessary pause for inspiration. This duration can be considered as a sort of physiological limit of phrase length.

The lungs can be thought of as a pair of balloons which are inflated for inspiration and deflated for expiration. These balloons are encased in a partly bony, partly muscular cylindrical cage. The ribs run around most of the cylinder, with a double muscular layer, the internal and external intercostal muscles, running between them, while the diaphragm,a large, dome-shaped muscle, forms its bottom. At the top, stems of the two balloons join, the larynx forms a valve at the top of the connecting, inverted Y-shaped tube, the trachea. In inspiration, the size of the lung cavity is increased by the contraction of the diaphragm and the lifting of the rib cage by the contraction of the external intercostals (and other muscles) (Ladefoged, 1967; Draper et al., 1958). When the lung cavity is increased in size, air flows into the lungs. Phonation occurs when the vocal folds are placed over, and set vibrating by, the expiratory air stream. To some extent, expiration is a passive phenomenon; the lungs are elastic in nature and, consequently, once inflated, will tend to deflate themselves. The functions of the muscles of respiration are organized in an ancillary way around this function. At the beginning of phonation after a deep inspiration, the muscles of inspiration can be used to break the outflow of air. At the end of expiration, the muscles of expiration (the internal intercostals and auxiliary muscles) can be used to prolong the breath-group, by squeezing more air out of the lungs, acting in opposition to the tissue forces which resist the deformation of the chest wall.

The point at which there is a changeover from the use of expiratory muscles to inspiratory muscles depends on the subglottal pressure maintained. At any time during the breath group, the internal intercostals can be used to produce momentary stress peaks (Ladefoged et al., 1958). Generally, subglottal pressure does not remain constant within a breath-group but falls at the end (Lieberman, 1967).

The effects of variation in subglottal pressure are two-fold. First, subglottal pressure affects the intensity of speech. Secondly, it can be shown that, everything else being equal, greater subglottal pressure will produce higher fundamental frequency. The mechanism for this process is described by Lieberman (1967). There are, then, three effects of speech respiration on the organization of speaking. First, the duration of expiration that can be sustained without an inspiration provides a physiological bound on phrase length; second, a respiratory mechanism permits the assignment of heavy stress; and third, an available subglottal mechanism accounts for the terminal pitch fall at the end of sentences.

The last two points have been the source of a good deal of discussion and controversy, which must be further discussed in connection with laryngeal mechanisms. For the present, let us just comment with respect to what we know about subglottal action itself.

First, Ladefoged has shown that, when a heavy stress is placed on one word in a sentence, there will be an accompanying increase in the action of the internal intercostal muscles. This should have the effect of producing the brief peaking in the subglottal pressure curve noted by Lieberman (1967) and Ohala (1970). There is no argument that this, in turn, will produce a peak in intensity in the acoustic speech output or that intensity rises, in turn, following Fry (1955), are one of the acoustic correlates of perceived stress. However, the subglottal peaks have been demonstrated only for very heavy, or contrastive, stress, which may not be a very central maneuver in ordinary running speech. Furthermore, there has been a general debate as to how large the effects of such peaks in subglottal pressure are on the fundamental frequency contour (Ohala, 1970), which is perhaps a more important correlate of perceived stress than is intensity (Fry, 1955).

The fall in subglottal pressure at the end of ordinary breath-groups is easily accounted for by the passive nature of the expiratory mechanism. If it is not compensated by the action of the laryngeal muscles, it will result in a fall in fundamental frequency, although here, again, there is argument about how much of the observed acoustic effect is subglottal in origin (see Ohala, 1970).

Laryngeal Mechanisms

Detailed description of the larynx is presented in Sawashima's chapter of this volume. Here, we will confine our attention to speech studies of the laryngeal muslces, particularly electromyographic studies. If this discussion is to be intelligible, however, it will be necessary to summarize some of the more general literature on laryngeal functions. The functions of the intrinsic muslces have been clarified by a long series of experiments by van den Berg and Tan (1959), in which they reconstructed the larynx of cadavers, duplicating normal air flow conditions, and modeled the effects of contraction of the various muscles. They describe the intrinsic muscles as having three general functions: tensing the vocal folds (the cricothyroid and vocalis), adducting the folds (the vocalis interarytenoids and the lateral cricoarytenoid muscles), and abducting the vocal folds. In general, the electromyographic studies of the action of these muscles in singing, from Faaborg-Anderson (1957) to the more recent work of Hirano et al. (1969) and Sawashima et al. (in press), have shown that the adductor muscles and the cricothyroid act together when pitch rises.

The action of the extrinsic laryngeal muscles is far less well understood. These are muscles which originate outside the larynx but insert on one of the laryngeal cartilages and hence influence the tension on the vocal folds and, consequently, pitch, indirectly. However, the direction of the influence will depend on interactions between the muscles above and below the larynx. For example, Hirano et al. (1967) and Hirano et al. (1969) have shown that one member of this group of muscles, the sternohyoid, is active at both high and low pitch extremes in singing. Sonninen (1956) has shown that persons who have had the sternohyoid, thyrohyoid, and omohyoid muscles sectioned for medical reasons typically have trouble singing high pitches after surgery. On the other hand, if the thyrohyoid, one of this group of muscles, is stimulated during surgery, pitch lowering may result, depending on the subject's head position.

Turning now to speech studies, the action of the laryngeal muscles has been studied only in a very few circumstances. There seems to be general agreement about what happens when a word is stressed in an English declarative sentence, resulting in an upward excursion of the pitch contour. As Hirano et al. (1969) have shown, there is, typically, a burst of activity in the cricothyroid, lateral cricoarytenoid, and vocalis muscles, accompanied by a peak in the fundamental frequency contour. Thus, for heavy stress, there is coordinated activity of laryngeal and subglottal systems to produce a peak in fundamental frequency and intensity, although Ohala (1970) believes that the effects of the subglottal system in producing the pitch rise are negligible. All three of these muscles can be shown to be active in the characteristic terminal rise for questions.

Considerable controversy surrounds the issue of pitch-lowering mechanisms in speech. The question can be subdivided into two substantive issues. First, is there an active pitch-lowering mechanism which is used for producing sudden downward pitch excursions in the course of an utterance? Second, is such a maneuver responsible for the fall in pitch at the end of declarative sentences?

English is perhaps not the ideal language for answering the first question, since, although Bolinger (1958) has demonstrated that sudden downward excursions in fundamental frequency are sometimes used to signal stress, this is not the common maneuver. Better examples are provided by Swedish, where a word accent distinction is signaled by variations in pitch contour. A model for Swedish word intonation has been suggested by Öhman (1967a). He suggests that the effects of Swedish grave and acute accents can be derived from a model that has positive sentence-intonation pulses and negative word-intonation pulses, which are differently timed for grave and acute accents and for different Scandinavian dialects. The model itself suggests, although it does not, of course, require, both a mechanism for pitch raising and an active mechanism for pitch lowering.

Two electromyographic investigations of Swedish word accents have been made. In the first, by Öhman et al. (1967), the cricothyroid and vocalis muscles were studied. Based on the results cited above, we would expect both cricothyroid and vocalis activity to be correlated with pitch rises. The most notable result of the study is an indication that there is a period of inhibition of cricothyroid activity which corresponds to the different times of application of the posited word intonation filter. Thus, there appears to be an inhibition of pitch-raising mechanisms that is correlated with pitch falls. No striking results were obtained for the vocalis.

A similar experiment was performed by Gårding et al. (1970), with probes in vocalis, cricothyroid, and sternohyoid muscles. Generally speaking, they find correlation between peaks in fundamental frequency and in the activity of cricothyroid and vocalis muscles. They find that "the sternohyoid activity shows no simple correlation with the pitch value."

Another language where sudden downward pitch excursions have a linguistic function is Tokyo Japanese, where the pitch level drops at the boundary following a vowel with an accent kernel mark. An electromyographic study of the cricothyroid, the lateral cricoarytenoid, and the sternohyoid was performed by Simada and Hirose (1970). In general, they find that there is a sharp fall in cricothyroid activity corresponding to the accent kernel. The activity pattern of the lateral cricoarytenoid is similar, although complicated by the participation of the muscle in voicing gestures. Most of their data do not show a clear correlation of the activity of the sternohyoid with the position of the accent kernel.

The experiments described above appear to indicate a passive pitch-lowering mechanism; that is, when pitch falls, the activity of the cricothyroid and the muscles that provide medial compression decreases. However, the only muscle that might lower pitch by increasing its activity which has thus far been examined, the sternohyoid, does not seem to indicate a clear pattern of correlation with pitch fall. It may be that other muscles are implicated.

The situation with respect to laryngeal adjustment at the termination of sentences is complicated by Lieberman's (1967) suggestion that the fundamental frequency fall is due to the speaker's failure to compensate at the larynx for falling subglottal pressure. This suggestion would be negated by either active or passive laryngeal adjustment. Speakers could either decrease the activity of the cricothyroid and its associated muscles at the end of a sentence or increase the activity of the sternohyoid, or some muscle with a similar activity pattern, at the ends of sentences. Inspection of cricothyroid records from Ohala's (1970) and our own (Harris et al., 1969) work does not reveal a characteristic fall-off in cricothryroid activity at sentence termination. Ohala (1970) has suggested that the sternohyoid has a tendency to be more active at sentence termination, but the picture is complicated, according to the later work of Ohala and Hirose (1970), by the tendency of the sternohyoid to participate in segmental gestures, such as jaw opening for open vowels. Again, other possible active muscles have been suggested for the pitch-lowering function.

In summary, then, the mechanism for raising pitch in speech has been demonstrated several times. When the cricothyroid contracts, pitch rises. Other muscles that contribute to medial compression of the vocal cords also contract, except as their function is complicated by their participation in voicing gestures. Pitch falls when the muscles that raise pitch relax or when subglottal pressure falls. In addition, an active mechanism for pitch lowering has been suggested, though not conclusively demonstrated. A further complication, not discussed here, is that actual pitch contour is influenced by the shape of the upper vocal tract (Flanagan and Landgraf, 1968).

The Upper Articulators

The third great group of articulatory muscles are those that are responsible for generating segmental phonemes. They, in turn, can be divided into a palate group, a tongue group, a group responsible for raising and lowering the jaw (which we will not describe), and those muscles of facial expression that act to mold the lips. For the linguist, it is probably confusing and unnecessary to supply an enormous list of muscle names and functions, cribbed chiefly from anatomy texts, which are not at present fleshed out by many physiological studies of muscle function in speech. However, sketching the anatomy of the oral region gives one a somewhat better idea how muscular organization limits vocal-tract shape. For the anatomically inclined, van Riper and Irwin (1958), some years ago, made some detailed guesses as to the muscle action involved in forming the vowels and consonants of American English. It is hard to see how these guesses could be improved on by anything except positive information.

The Tongue.

Let us begin with the tongue. It has two great muscle systems--extrinsic muscles, which connect the tongue to another structure, and intrinsic muscles, whose fibers run entirely within the tongue body.

The tongue can be moved forward by the genioglossus muscle, a fan-shaped muscle whose fibers make up a great part of the core of the tongue. It can be moved up and back by the styloglossus, which runs from the sides and back of the tongue to the styloid process of the skull just behind the ear. It can be moved down and back by the hyoglossus, which runs from the sides of the tongue to the horns of the hyoid bone, a horseshoe shaped bone which forms an underpinning for the tongue. In addition to direct pull in these three directions, the tongue may be pulled in the same directions by muscles connecting the hyoid bone to other structures.

The intrinsic muscles of the tongue are named for their fiber directions: the transverse and vertical muscles, which probably groove and flatten the tongue, and the inferior and superior longitudinals, which probably curve the surface up and down. It is generally believed that these muscle fibers act together to shape the tongue tip.

There has been very little electromyographic work on the tongue. The technical problems involved are quite difficult; in particular, the question of what muscle is being examined when a probe is inserted makes assignment of function very difficult. The recent development of a flexible wire electrode (Hirano and Ohala, 1969) should greatly facilitate work of this type in the future. However, our knowledge of the movements of the tongue and its associated stuctures has been greatly enlarged by two recent cineradiographic studies by Houde (1967) and Perkell (1969). Most of their discussion is concerned with general models of the articulatory process; two specific hypotheses might perhaps better be described here.

The first is the observation by Perkell that the vowels of American English conform to the tense/lax description proposed originally by Jakobson et al. (1963) and discussed at greater length by Chomsky and Halle (1968). The latter state: "Tense sounds are produced with a deliberate, accurate, maximally distinct gesture that involves considerable muscular effort; non-tense sounds are produced rapidly and somewhat indistinctly" (p. 321). Perkell believes that he observes such differences in vocal tract adjustments of pairs such as /i/ and /I/. He cites the work of MacNeilage and Sholes (1964) on the electromyographic activity of the tongue to provide further support for this notion; they show that tense vowels have generally higher EMG voltage levels than their lax counterparts.

Indeed, the MacNeilage and Sholes data show duration differences between /i/ and /I/ quite clearly, as one might expect. However, if one examines their data, the ranking of total EMG activity over all the vowels does not conform to a tense/lax classification. Furthermore, the MacNeilage and Sholes data are taken from a very restricted set of sampling points and, consequently, cannot be used for inferences about activity in the tongue musculature as a whole.

Perkell suggests also that the extrinsic muscles of the tongue are used for vowel production, while both intrinsic and extrinsic muscles are used for consonants. He points out that the tongue behaves like a semirigid body in the production of vowels: its shape is more or less constant, and it is moved into target position by the extrinsics. The same observation of the constancy of the tongue shape is made by Houde (1967). At present, it is not possible to assign this result to a specific muscle set. Perkell believes that the intrinsic tongue muscles are particularly implicated in consonant production. It seems reasonable that the intrinsic muscles of the tongue should be responsible for the complicated shaping of the tip in producing apical consonants. It is less easy to see how they could be primarily responsible for /k/ closure. In this case, the extrinsic muscles may well pull the body of the tongue up and back, without changing its shape in detail. Of course, any such hypothesis must wait on the development of better techniques for electromyographic study of the tongue muscles. It would be interesting to evaluate the hypothesis with a wider range of back tongue consonants than Perkell uses.

The Velopharyngeal Closure System. Velopharyngeal closure has probably been more extensively studied by physiological techniques than has any other part of the speech mechanism. The reasons for this have to do with the clinical problems of its repair.

The action of the palate in making velopharyngeal closure has been studied by both X-ray and electromyographic techniques. X-ray studies are summarized by Bjork (1961) and Nylen (1961). Electromyographic studies are summarized and extended in a recent monograph by Fritzell (1969).

Briefly, velopharyngeal closure is accomplished by elevating the soft palate to block the nasal passageway when an oral sound is produced. The chief active agent would appear to be the levator palantini muscle, which makes up the bulk of the soft palate (Fritzell, 1969). Older accounts (Bloomer, 1953) suggest that velopharyngeal closure is accomplished by a kind of "purse-string" action--that is, the sides of the pharyngeal port come in as the palate elevates, while the posterior pharyngeal wall comes forward. More recent X-ray studies suggest that this purse-string action of the posterior pharyngeal wall is not common in normal speakers (Hagerty et al., 1958). Furthermore, one EMG study (Harris and Schvey, 1962) has shown that, while the muscles of the upper pharyngeal wall are quite active in speech, this activity is not well correlated with the action of the palate. However, it can be shown to be a mechanism occasionally used in effecting velopharyngeal closure in persons with insufficient tissue to make the closure with the velum alone. This difference in mechanism is interesting from the point of view of the general theory of phoneme formation, since it suggests that a wide range of individual differences in gesture are possible for the achievement of the same result.

A second interesting question about articulatory dynamics has been raised in discussion of the nature of velopharyngeal opening after closure. The situation is very much like that for pitch lowering. It is not clear whether velar opening is under active muscular control or whether it is accomplished by gravity, in addition to the relaxation of the muscles of velar closure. Fritzell (1969) has some preliminary data which suggest that there is active contraction of muscles which lower the velum when a nasal follows an oral consonant--that is, that there is active velar lowering--but further research is clearly necessary on this point.

Another question that has been studied is how many different degrees of velar activity are necessary in speech--i.e., whether there is a partial reorganization of the oralization gesture which depends on other aspects of the phoneme. Fritzell (1969) presents clear evidence which supports the earlie work of Lubker (1968) by showing that the extent of palate movement for high vowels is greater than for low vowels--that is, that the oralization gesture is reorganized. We will discuss this result further below.

Facial Muscles. The third great group of muscles involved in articulation are those that shape the lips--also known as the muscles of facial expression, for obvious reasons. Generally speaking, the muscles shaping the mouth can be conceived as falling into two functional groups: the orbicularis oris fibers, which form a heavy band around the mouth opening and act in sphincter fashion to round and purse the lips, and a series of muscle bundles that insert radially into the oris bundle from various directions. These bundles have an action that depends on their insertion--to pull the upper lip up, to pull the lower lip down, and to spread the lips laterally. In general, all the labial consonants of English show an implosion peak of orbicularis oris activity and an explosion peak in the muscles that withdraw the lips.

In general, it has been found that the size of the closure peak for English /p/, /b/, and /m/ appears to be the same regardless of the following vowel or which of the three consonants is being produced (Harris et al., 1965; Fromkin, 1966; Tatham and Morton, 1968). One study (Öhman, 1967c), however, has found a small difference between closure peak sizes for /p/, /b/, and /m/ (using a Swedish speaker). The size of the explosion peak in the muscles that withdraw the lips varies systematically with the following vowel (MacNeila and deClerk, 1969). All these results will be discussed below in connection with theories about motor organization in speech.

THE ORGANIZATION OF SPEECH

Both MacNeilage (1970) and Öhman (1967c) have concluded that a description of articulatory dynamics should have the phoneme as its basic unit, in spite of the elusiveness of its physical manifestations. As we have seen, there is nothing necessary about this view of the articulatory process. Mattingly's chapter of this volume summarizes some of the problems of this point of view as a scheme for speech synthesis. In this section, we will try to indicate the problems in view of what is now known about physiological articulatory phonetics.

Allophonic Variants

We have already discussed MacNeilage's (1970) suggestion that if the phoneme is to be considered as a basic unit, it must be stored in the nervous system as a positional target rather than as a movement. He suggests that the motor system is controlled by the results of an internal specification of certain spatial targets. This accounts, presumably, for the results of MacNeilage and deClerk (1969), who showed that, in the production of CVC monosyllables, the lingual electromyographic signals for the vowel are regularly conditioned by the preceding consonant. Fromkin (1966) and Öhman (1967c) have shown similar results for the muscles controlling the lips. For example, Fromkin showed that the amount of activity associated with the rounded vowel /u/ is greater in the context /dud/ than in the context /bub/, presumably because in the latter context the lips are closer to the rounding target for the vowel than in the former.

Given that the articulators tend to adopt a fixed position for a given phoneme, it is still in question whether stress allophones of the phonemes are commonly preserved--that is, whether phonemes are more strongly articulated in positions of heavy stress. The evidence for stress allophones is presented in connection with the section "undershoot," below.

The question of stress comes up in connection with the "feature" argument. Jakobson et al. (1963) proposed that certain consonants and vowels differ, pairwise, from each other in a "tense/lax" dimension. The "tense" member of the pair was supposed to be a more forceful articulation than the "lax." This distinction has already been discussed relative to the vowels. The same distinction, in their system, is meant to be the primary distinction between "voiced" and "voiceless" consonants. In particular, /p/ is meant to be distinguished from /b/ primarily along a tension dimension. Based on this assumption, one would expect that the orbicularis oris contraction for the /p/ closure would be more forceful than for the /b/ closure. This expectation has been examined in four studies (Harris et al., 1965; Fromkin,1966; Tatham and Morton,1968; Öhman, 1967c). In the first three, differences between /p/ and /b/ have been found to be insignificant, while a very small peak-contraction size difference was found in the fourth.

Muscular tension differences have been suggested as an explanation for /p/ - /b/ distinctions in quite another way. Chomsky and Halle (1968), followed by Perkell (1969), have suggested that the observed difference in upper pharyngeal tract size is a passive response to the fact that the upper tract muscles are "tenser" and hence hold the tract walls more rigid for the tense consonant. The larger pharynx size for voiced consonants is confirmed by Kent and Moll (1969). They feel, however, that the size adjustment is under active muscular control: the extrinsic muscles act on the larynx to lower it and hence increase the size of the upper vocal tract. The details of this extremely complicated physiogical argument are discussed in Lisker and Abramson (1971) and in Lisker's chapter of this volume. The weight of the evidence seems to suggest that mechanisms other than generalized tensing are responsible for the perceived voiced/voiceless distinction.

It is hard for me to imagine the detailed workings of a system in which both stress differences and phonemic differences were maintained by the same general physiological mechanism. My own present view is that the voiced/voiceless distinction is carried by the timing of glottal adjustment, while the peripheral articulation of cognate pairs is the same. This implies that place and voicing features can be spatially segregated, and it also implies general articulatory separation of features.

A preliminary study (Harris et al., 1962) suggested that the oralization gesture might be the same for voiced and voiceless consonants and, thus, similarly preserves the independent organization of features. However, Lubker (1968) has shown that, for vowels, the oralization gesture, that is, the size of the velopharyngeal closure gesture, is not independent of the vowel height. Therefore, we cannot assume a general orthogonal articulatory organization of features.

If we assume that phonemes are stored as constant targets, we must develop ways of specifying and explaining the obvious failure of the articulators to move from one invariant target to the next in running speech. Using Öhman's terminology, three mechanisms have been proposed to account for allophonic variation--reorganization, undershoot, and coarticulation.

Reorganization

"Reorganization" is a grab-bag term intended to cover cases where allophones have different articulations in a fundamental sense--i.e., there is a context-dependent change in feature specifications. The two examples given by Öhman, devoicing of final voiced consonants in German and Russian and quality alternations of vowels under vowel harmony, are language specific; at present, no general theory of speech production accounts for the particular circumstances in which reorganization should occur. However, the other two mechanisms are presumed to have universal application, and some general statements have been made about them.

Undershoot

Undershoot results when "an incomplete articulatory gesture is interrupted by a neural command that brings about the next gesture of the utterance." The key experiment for the demonstrations of undershoot was performed by Lindblom (1963) on vowel reduction. He had subjects produce CVC syllables in a sentence frame such that the stress on the syllable was varied. As a result, the duration of the vowel varied; the shorter the vowel, the further from a target frequency the vowel formants fell. Lindblom was able to show that the failure of target attainment for the vowel could be predicted from a simple model. In the model, there is presumed to be a simple activating command for each of the three phoneme elements of the syllable. Due to such factors as the inertia of the articulatory structures, there is a time delay between the arrival of the command and movement completion. If successive commands arrive fast enough, the moving articulators will not attain target position. Lindblom also showed that the acoustic effects of increased speaking rate on vowel target attainment are the same as reduced stress.

Lindblom's model is based on the assumption, mentioned above, that stress does not affect the magnitude of vowel commands but merely their timing. Although he does not examine the question, it seems likely that he intends the undershoot mechanism to apply to consonants as well as vowels.

No electromyographic studies have been made, so far as I know, to test Lindblom's undershoot model directly. However, the prediction, by extrapolation, to the EMG level should be that, under conditions of varying stress, the EMG signals associated with phonemes remain constant but the spacing between them is altered. We do not have evidence about the behavior of vowels under varying stress, but we have performed a relevant experiment on consonants (Harris et al., 1968). When we compared the amplitude of electromyographic signals (orbicularis muscle) in words that were heavily emphasized in sentences with the same words unemphasized, we found that the stressed amplitude was some 10-20 percent greater. However, changes in lexical stress were not accompanied by amplitude changes. Fromkin in an earlier study (1966) has shown a larger amplitude in the orbicularis signals for initial /b/ closure in words than for terminal /b/ closure, although a related study (Harris et al., 1965) did not show a similar effect. From what we know about the way in which the EMG signal is related to the resulting articulatory motion, a larger amplitude of EMG signal is translated into a faster acoustic transition and/or a more extreme articulatory position. Lindblom's model implies that this does not happen. It seems important to perform a direct EMG-analog of the Lindblom experiment and also to check the limits of the size effect. On the basis of present knowledge, a size change effect sometimes does occur. However, this may or may not represent an extreme maneuver, outside the range of the normal stress variation studied by Lindblom. Certainly, a system with signals of variable spacing but constant size is somewhat simpler to analyze than one in which signal size varies with stress and tempo of articulation.

Anticipatory Coarticulation

It has long been recognized that certain characteristics of a phone are likely to be anticipated in running articulation. For example, if a vowel is rounded, signs of the vowel rounding will occur in the articulatory gesture of the preceding consonant. Of course, this anticipation will act to produce different target configurations for the same phone in front of other different phones, with consequent acoustic and electromyographic effects.

This phenomenon appears to arise in a sequence of phones, the second of which is marked with respect to a characteristic that is unmarked in the first. For example, in English, the vowels are marked with respect to rounding, while many consonants are not. Therefore, such a CV sequence will show vowel rounding during the consonant. Similar types of anticipation during consonant production of aspects of the vowel are shown by MacNeilage and deClerk (1969). The determinants of the length of an anticipatory sequence are not known. Kozhevnikov and Chistovich (1965) have suggested that coarticulation boundaries act to delimit the syllable. In their model, the commands for a syllable are specified simultaneously with the start of the first command for which the commands are noncompleting, or at the syllable boundary. Consequently, there is variable coarticulation within the syllable but minimum coarticulation between syllables. Two recent experiments, by Daniloff and Moll (1968) and by Amerman et al. (1970), appear to contradict this view. Using measurements of X-ray film, they were able to show that coarticulation may extend over several phoneme units, even when word or syllable boundaries occur in the sequence. Some acoustic measurements made by Öhman (1966) of VCV sequences similarly show anticipatory coarticulation of the second vowel across the consonant, i.e., across the syllable boundary. He suggests that consonant gestures are somehow superimposed on slowly varying vowel targets. In all these examples, characteristics of the vowel are anticipated through a string of consonants. However, Amerman et al. cite an as yet (1970) incompletely reported experiment which shows anticipation of nasalization through a vowel sequence, again without regard to syllable boundaries.

No present model of articulatory behavior seems adequate to describe the circumstances and extent of anticipatory coarticulation. Syllable boundaries do not seem, in fact, to play the essential role that Kozhevnikov and Chistovich suggest. At least on the basis of present information, there does not seem to be the distinction of function between vowels and consonants in coarticulation that Öhman describes. Henke (1967) posits that each phoneme can be specified as a bundle of features, although each phoneme is not specified with regard to each feature. In the production of any sequence of phonemes, a high-level scan looks ahead from one positive value of a feature over a series of neutral values to the next specific positive value of the feature. However, Henke's description does not specify the time course of anticipatory coarticulation in detail or, indeed, what characteristics of intervening phones affect it.

In short, then, although we are in a position now to specify some apparently general mechanisms that interfere with attainment of constant target, the scope of these mechanisms is not presently understood. One further question is how even approximate target maintenance is achieved. The usual explanation proposed (by MacNeilage, for example) is that the speaker keeps tract of articulatory position by feedback of some sort.

The Role of Feedback

Two types of feedback are clearly used at some stage in language development. The first of these is acoustic feedback. Acoustic feedback is clearly necessary for adequate speech development, as is demonstrated by the common failure of deaf children to speak intelligibly. Indeed, it has been suggested by Whetnall and Fry (1964) that children whose hearing is seriously impaired will speak normally if the remaining hearing is efficiently used. They suggest that the essential prerequisite to intelligible speech is an association between an articulation and some form of acoustic image, even if the acoustic image is seriously distorted.

Acoustic feedback has a somewhat more equivocal role in speech that has already developed. Noise masking of speech has little or no effect on speech intelligibility in adult speakers (see, for example, Ringel and Steer, 1963). Anecdotal experience suggests that traumatic deafness, in adulthood, does not cause immediate degradation of speech quality. On the other hand, continuous acoustic monitoring must have some role in speech maintenance; the speech of deafened adults does apparently deteriorate eventually, although this phenomenon has not been adequately studied. Furthermore, delayed auditory feedback has devastating effects on speech. In the delayed auditory feedback situation, the speaker is fitted with earphones and hears his own speech after a delay; the appropriate delay will cause slowing of speech, distortion of production, and stuttering (Fairbanks and Guttman, 1958). In summary, then, continuous auditory monitoring appears to be unnecessary for speech production, but maintenance of normal articulation cannot survive serious distortion or deprivation of auditory feedback.

The second type of feedback believed to be important for the preservation of the integrity of speech is the feedback from the oral articulatory structures. This feedback may be one of two types, although they are not clearly distinguished in much of the literature. First, important feedback information may be conveyed by sensations arising from contact between oral structures, for example, the contact between the hard palate and the tongue tip. It is well known that the whole oral area is well endowed with sensory receptors for touch and pressure; Ringel (1970) has recently reviewed the literature on sensitivity of the oral region.

A second type of feedback that may be important in phoneme target maintenance is proprioception, defined by Scott (1970) as "the sensation which results in knowledge...of articulator position and movement which is traditionally believed to be mediated by muscle and jaw receptors," that is, information about the state of the muscle itself. Appropriate muscle receptors have recently been located in the laryngeal muscles by Baken (1969). The evidence for appropriate receptors in the supralaryngeal muscles is summarized by Scott.

Assuming that such a mechanism is present, there is still the question of whether it is used in moment-to-moment monitoring of articulator position. Three types of evidence have been offered.

First, there is some clinical evidence that persons with reduced somesthetic perception will have deficient speech. MacNeilage et al. (1967) studied a patient with a grossly disturbed somesthetic system whose speech was almost wholly unintelligible. Cases of this sort, however, are difficult to classify. MacDonald and Aungst (1970) studied another patient who might be described as having the same pattern of deficit but whose speech was, to ear at least, perfectly normal.

Perhaps more convincing evidence comes from experiments on oral stereognosis. It has been shown (Shelton et al., 1967) that normal adults can recognize three-dimensional objects which are placed in their mouths. A recent study (Ringel et al., 1970) has shown that children with articulatory deficiencies are significantly less proficient at the oral stereognosis task. However, the oral stereognosis task is rather complicated, and its association with other types of performance has not yet been thoroughly studied.

A second type of evidence for the importance of sensory feedback in speech maintenance comes from examination of the effects of anesthetization of parts of the upper vocal tract. The classic study was performed by Ringel and Steer (1963) and has recently been repeated with some technical improvement by Scott (1970). In Scott's study, injection of topical anesthetic was believed to eliminate sensation from surface receptors in the tongue, palate, teeth, lips, and oral mucosa. Since the innervation of the kinesthetic receptors in the oral region is not completely understood, it is not known whether feedback from them was also blocked.

The chief effects of the multiple deprivation appear to be a reduction of the ability to refine labial and tongue blade and tip articulations, although speakers are able to maintain fairly intelligible speech. The interpretation of the findings as to feedback is, however, open to some question.

A recent experiment by Borden (reported by Harris, 1970) shows that, when the sensory block from the tongue is performed as in the experiments cited above, two of the muscles that may control positioning of the tongue tip can be paralyzed by spread of the injected anesthetic. It is not clear, then, the extent to which the results obtained are due to surface sensory deprivation, kinesthetic sensory deprivation, or motor paralysis.

The third type of evidence is from experiments where a significant correlation has been shown between successive aspects of the articulation of a syllable (MacNeilage, 1969; Ohala, 1970; Kozhevnikov and Chistovich, 1965). For example, the latter authors show a positive correlation between maximum jaw opening and the velocity of jaw opening and closing in the production of CVC monosyllables.

Kozhevnikov and Chistovich (1965) concluded, from the results of an earlier experiment: "Lowering of the...lower jaw must lead to a reflexive increase in excitability of the centers of antagonistic muscles, while excitability must increase all the more with a greater drop of the...lower jaw." Ohala felt that "this...gives fairly good evidence of the presence and use of short-term feedback to make quick adjustments of articulator movement in speech" (p. 41). Again, the argument seems to be a little difficult. First, acoustic feedback, as well as kinesthetic feedback, was available to the subject, and second, one could argue that the results are accounted for by some general fluctuation in excitability affecting several successive gestures, rather than an actual feedback of one upon the next. One must agree with MacNeilage that the evidence for closed-loop control of speech is suggestive rather than definitive.

PHYSIOLOGICAL REPRESENTATION OF THE PHONEME

On the basis of our present knowledge, we appear to have failed to find a simple, absolutely invariant correlate of the phoneme at the peripheral levels thus far investigated.

The inertia of articulatory structures clearly limits the extent to which invariant positions can be attained on a phoneme-by-phoneme basis. Furthermore, there is considerable temporal uncertainty as to the time of application of parts of commands corresponding to a single perceptual phoneme entity. We should perhaps have anticipated the failure to find shape invariance, since attempts to find acoustic invariance, even for single speakers, have failed in the past. Clearly, if a listener extracts a string of invariant phonemes from the acoustic flow, he must do so on the basis of running perception of units that are larger than phoneme size, although strict syllable-by-syllable chunking does not seem to work very well either. Since there is no phoneme-by-phoneme progression in articulation, there is no need for precise closed-loop feedback control of articulatory position, nor is the evidence for such feedback very strong. Monitoring appears to be necessary only in some more general sense.

On the other hand, we do not appear to be very far away from appreciating the proper size of physiological segments. No one has proposed that articulatory context dependencies extend very far. If the syllable, as traditionally defined, is not the proper size unit, it is probably quite close to it. Only by careful, continuing study of articulatory coarticulation can we refine our knowledge of the field size over which perceptual scanning operates.

REFERENCES

Amerman, James D., Raymond Daniloff, and Kenneth Moll. 1970. Lip and jaw coarticulation for the phoneme /ae/. JSHR. 13.147-161.

Baken, Ronald J. 1969. Distribution of neuromuscular spindles in intrinsic muscles of a human larynx. Unpublished Ph.D. dissertation, Columbia University.

Basmajian, J.V. 1962. Muscles alive: Their functions revealed by electromyography. Baltimore: Williams and Wilkins.

Bjork, L. 1961. Velopharyngeal function in connection speech. Acta Radiologica Suppl. 202.

Bloomer, Harlan H. 1953. Observations on palatopharyngeal movement in speech and deglutition. JSHD. 18.230-246.

Bolinger, Dwight L. 1958. A theory of pitch accent in English. Word 14.109-149.

Chomsky, Noam and Morris Halle. 1968. The Sound Pattern of English. New York: Harper & Row.

Cooper, Franklin S. 1965. Research techniques and instrumentation: EMG. Proceedings of the Conference on Communicative Problems in Cleft Palate. ASHA Reports 1.53-168.

Daniloff, Raymond G. and Kenneth Moll. 1968. Coarticulation of lip-rounding. JSHR. 11.707-721.

Draper, M.H., Peter Ladefoged, and D. Whitteridge. 1958. Respiratory muscles in speech. JSHR. 2.16-27.

Faaborg-Andersen, K. 1957. Electromyographic investigation of intrinsic laryngeal muscles in humans. Acta Physiologica Scandinavica 41, Suppl. 140.

Fairbanks, Grant and Newman Guttman. 1958. Effects of delayed auditory feedback upon articulation. JSHR. 1.12-22.

Fant, C. Gunnar M. 1960. Acoustic Theory of Speech Production. The Hague: Mouton.

Flanagan, James L. and Lorinda L. Landgraf. 1968. IEE Transactions on Audio and Electroacoustics 16.57-64.

Fritzell, Bjorn. 1969. The velopharyngeal muscles in speech: An electromyographic and cineradiographic study. Acta Otolaryngologica Suppl. 250. 1-81.

Fromkin, Victoria A. 1966. Neuromuscular specification of linguistic units. L & S. 9.170-199.

Fry, Dennis B. 1955. Duration and intensity as physical correlates of linguistic stress. JAcS. 27.765-768.

Gårding, Eva, Osamu Fujimura, and Hajime Hirose. 1970. Laryngeal control of Swedish word tone--a preliminary report on an EMG study. Annual Bulletin, Research Institute of Logopedics and Phoniatrics 4.45-54. Tokyo: University of Tokyo.

Hagerty, Robert F., Milton J. Hill, Harold S. Pettit, and John J. Kane. 1958. Posterior pharyngeal wall movement in normals. JSHR. 1.203-210.

Harris, Katherine S., Malcolm M. Schvey, and Gloria Lysaught. 1962. Component gestures in the production of oral and nasal labial stops. JAcS. 34.743(A).

Harris, Katherine S., Gloria F. Lysaught, and Malcolm M. Schvey. 1965. Some aspects of the production of oral and nasal labial stops. L & S. 8.135-147.

Harris, Katherine S., Thomas Gay, George N. Sholes, and Philip Lieberman. 1968. Some stress effects on the electromyographic measures of consonant articulations. Reports of the 6th International Congress on Acoustics. Y. Kohasi, Ed. Tokyo: International Council of Scientific Unions.

Harris, Katherine S., Thomas Gay, Philip Lieberman, and George N. Sholes. 1969. The function of muscles in control of stress and intonation. Haskins Laboratories Status Report on Speech Research 19/20.127-38. New York: Haskins Laboratories.

Harris, Katherine S. In press. Physiologic measures of speech movements: EMG and fiberoptic studies. State of the Art Conference: Speech and the Dento-Facial Complex, 1970. Washington, D.C.: American Speech and Hearing Association.

Henke, William. 1967. Preliminaries to speech synthesis based on an articulatory model. Conference preprints, 1967 Conference on Speech Communication and Processing, 170-177. Bedford, Mass.: Air Force Cambridge Research Laboratories.

Hirano, Minoru, Yasuo Koike, and Hans von Leden. 1967. The sternohyoid muscle during phonation. Acta Otolaryngologica 64.500-507.

Hirano, Minoru and Timothy Smith. 1967. Electromyographic study of tongue function in speech: A preliminary report. Working Papers in Phonetics 7.45-56. Los Angeles: UCLA Phonetics Laboratory.

Hirano, Minoru and John Ohala. 1969. Use of hooked-wire electrodes for electromyography of the intrinsic laryngeal muscles. JSHR. 12.362-373.

Hirano, Minoru, John Ohala, and William Vennard. 1969. The function of laryngeal muscles in regulating fundamental frequency and intensity of phonation. JSHR. 12.616-628.

Houde, Robert A. 1967. A study of tongue body motion during selected speech sounds. Speech Communications Research Laboratory monogr. No. 2. Santa Barbara: Speech Communications Research Laboratory.

Jakobson, Roman, C. Gunnar M. Fant, and Morris Halle. 1963. Preliminaries to speech analysis. Cambridge, Mass.: M.I.T. Press. Originally published as Technical Report No. 13, Acoustics Laboratory, M.I.T.

Kent, Raymond D. and Kenneth L. Moll. 1969. Vocal-tract characteristics of the stop cognates. JAcS. 46.1549-1555.

Kozhevnikov, V.A., and Ludmila A. Chistovich. 1965. Rech'Artikuliatsia i Vospriiatie. Trans. as Speech: Articulation and Perception. 1966. Washington, D.C.: Joint Publications Research Service.

Ladefoged, Peter, M.H. Draper, and D. Whitteridge. 1958. Syllables and stress. Miscellanea phonetica 3.1-14.

Ladefoged, Peter. 1967. Three Areas of Experimental Phonetics. London: Oxford University Press.

Liberman, Alvin M., Franklin S. Cooper, Donald P. Shankweiler, and Michael Studdert-Kennedy. 1967. Perception of the speech code. Psychological Review 74.431-461.

Lieberman, Philip. 1967. Intonation, Perception and Language. Cambridge, Mass.: M.I.T. Press.

Lindblom, Bjorn. 1963. Spectrographic study of vowel reduction. JAcS. 35. 1773-1781.

Lisker, Leigh and Arthur S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. Word 20.384-422.

Lisker, Leigh and Arthur S. Abramson. 1971. Distinctive features and laryngeal control. Language 47. To appear.

Lubker, James F. 1968. An electromyographic-cinefluorographic investigation of velar function during speech production. Cleft Palate 5.1-18.

MacDonald, Eugene and Lester Aungst. 1970. Apparent independence of oral sensory functions and articulatory proficiency. Symposium on Oral Sensation and Perception. James Bosma, Ed. 391-397. Springfield, Ill.: Thomas.

MacNeilage, Peter F. and George N. Sholes. 1964. An electromyographic study of the tongue during vowel production. JSHR. 7.209-232.

MacNeilage, Peter F., Thomas P. Rootes, and Richard A. Chase. 1967. Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. JSHR. 10.449-467.

MacNeilage, Peter F. and Joseph L. deClerk. 1969. On the motor control of coarticulation in CVC monosyllables. JAcS. 45.1217-1233.

MacNeilage, Peter F. 1970. Motor control of serial ordering of speech. Psychological Review 77.182-195.

Nylen, Bengt O. 1961. Cleft palate and speech. Acta Radiologica Suppl. 203.

Ohala, John and Hajime Hirose. 1970. The function of the sternohyoid muscle in speech. Annual Bulletin, Research Institute of Logopedics and Phoniatrics 4.41-44. Tokyo: University of Tokyo.

Ohala, John. 1970. Aspects of the control and production of speech. Working Papers in Phonetics 15. Los Angeles: UCLA Phonetics Laboratory.

Öhman, Sven E.G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. JAcS. 39.151-168.

Öhman, Sven E.G. 1967a. Word and sentence intonation: A quantitative model. Speech Transmission Laboratory Quarterly Progress and Status Report 2-3. 20-55. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology.

Öhman, Sven E.G. 1967b. Numerical model of coarticulation. JAcS. 41.310-320.

Öhman, Sven E.G. 1967c. Peripheral motor commands in labial articulation. Speech Transmission Laboratory Quarterly Progress and Status Report 4.1-29. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology.

Öhman, Sven E.G., A. Martensson, R. Leandersson, and A. Persson. 1967. Crico-thyroid and vocalis muscle activity in the production of Swedish tonal accents: A pilot study. Speech Transmission Laboratory Quarterly Progress and Status Report 2-3.55-57. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology.

Perkell, Joseph S. 1969. Physiology of speech production: Results and implications of a quantitative cineradiographic study. Cambridge, Mass: M.I.T. Press.

Ringel, Robert L. and Max D. Steer. 1963. Some effects of tactile and auditory alterations on speech output. JSHR. 6.369-378.

Ringel, Robert L., Arthur S. House, Kenneth W. Burk, John P. Dolinsky, and Cheryl M. Scott. 1970. Some relations between orosensory discrimination and articulatory aspects of speech production. JSHD. 35.3-11.

Ringel, Robert L. In press. Oral sensation and perception: A selective review. State of the Art Conference: Speech and the Dento-Facial Complex, 1970. Washington, D.C.: American Speech and Hearing Association.

Sawashima, Masayuki, Thomas Gay, and Katherine S. Harris. In press. An electromyographic study of the pitch and intensity control mechanisms of the larynx. JSHR.

Scott, Cheryl M. 1970. A phonetic analysis of the effects of oral sensory deprivation. Unpublished Ph.D. dissertation, Purdue University.

Shelton, Ralph L., William B. Arndt, Jr., and John Jones Hetherington. 1967. Testing oral stereognosis. Symposium on oral sensation and perception. James Bosma, Ed. 221-243. Springfield, Ill.: Thomas.

Simada, Zyun'ichi and Hajime Hirose. 1970. The function of the laryngeal muscles in respect to the word accent distinction. Annual Bulletin, Research Institute of Logopedics and Phoniatrics 4.27-40. Tokyo: University of Tokyo.

Sonninen, Aato. 1956. The role of the external laryngeal muscles in length-adjustment of the vocal cords in singing. Acta Otolaryngologica Suppl. 130.

Stevens, Kenneth N. 1970. Perception of phonetic segements: Evidence from phonology, acoustics and psychoacoustics. In Perception of Language. Paul Kjeldergaard, Ed. Columbus, Ohio: Chas. E. Merrill.

Tatham, Marcel Q.A. and Katherine Morton. 1968. Some electromyography data towards a model of speech production. Occasional Papers 1.1-59. Colchester: University of Essex Language Centre.

Van den Berg, Janwillem and T.S. Tan. 1959. Results of experiments with human larynxes. Practica Oto-Rhino Laryngologica 21.425-450.

Van Riper, Charles and John V. Irwin. 1958. Voice and Articulation. Englewood Cliffs, N.J.: Prentice-Hall.

Whetnall, Edith and Dennis B. Fry. 1964. The Deaf Child. London: Heinemann.

Zemlin, Willard. 1968. Speech and Hearing Science: Anatomy and Physiology. Englewood Cliffs, N.J.: Prentice-Hall.

Laryngeal Research in Experimental Phonetics[*]

Masayuki Sawashima[+]
Haskins Laboratories, New Haven

RESEARCH ON SOME BASIC ASPECTS OF THE LARYNX

The origin of the larynx is seen in a simple sphincter mechanism at the entrance of the airway of the lungfish. Along the evolutionary path, the larynx has been modified to a complicated structure with several cartilages and muscles. Although the original sphincteric use remains unchanged, some additional functions have been developed for the larynx along with its structural modifications and with ecological variation of animals. The relationships among anatomy, physiology, and ecology in the development of the larynx has been well described by Negus (1962) in his famous book The Comparative Anatomy and Physiology of the Larynx. Among those additional functions, the most important one, especially for the human larynx, is that of phonation. It is true that the well-developed ability to produce voice and speech is exclusive to mankind; however, this does not necessarily mean that the human larynx, efficient sound generator that it is, is especially modified for this activity. As in other animals, protective closure of the airway against foreign bodies is also a major function of the human larynx.

Many experimental studies have concentrated on features of the larynx as a vocal organ; however, these features cannot be separated from those contributing to the protective sphincteric function. Current progress in experimental phonetics requires a more profound understanding of basic laryngeal features. In this section, recent research will be outlined.

Structure and Function of the Laryngeal Muscles

The classical concept of the structure of the vocalis muscle was that the muscle fibers ran parallel to the vocal ligament between the thyroid cartilage and the arytenoid cartilage. A report by Goerttler (1950) has stimulated further examination of the fiber arrangement in the vocalis muscle. According to Goerttler, contrary to the classical position, the muscle fibers are divided into two groups (anterior and posterior), both of them inserting into the vocal ligament. The anterior division of the vocalis muscle (portio thyreovocalis) arises from the posterior surface of the thyroid cartilage and runs in postero-medial direction; the posterior division (portio aryvocalis) arises from the muscular process of the arytenoid cartilage and runs in antero-medial direction. Muscle fibers of the two groups cross each other in their course to the vocal ligament. This fiber arrangement suggests the possibility of opening the membranous portion of the glottis by contraction of the vocalis muscle, which would seem to support Husson's neurochronaxic theory of vocal-fold vibration.

---

[*] Chapter prepared for Current Trends in Linguistics, Vol. XII, Thomas A. Sebeok, Ed. (The Hague: Mouton).

[+] On leave from the University of Tokyo.

Goerttler's study has been criticized by Wustrow (1952) and many other investigators. According to Wustrow, the main fibers of the vocalis muscle run parallel to the vocal ligament. There are a few muscle fibers that arise from the arytenoid and insert into the vocal ligament (Goerttler's portio aryvocalis), but they are of no importance in controlling glottal opening. There are no muscle fibers inserting into the vocal ligament from the thyroid cartilage (Goerttler's portio thyreovocalis). Wustrow also divided the fibers of the vocalis muscle into two groups. Both arise from the thyroid cartilage, one superior to the other, and insert into the arytenoid cartilage; the superior group runs to the vocal process of the arytenoid cartilage, while the inferior group reaches a point lateral to the vocal process. Thus, the two muscle bundles are parallel to each other when the arytenoid cartilage is abducted but twisted when the cartilage is adducted. Wustrow's findings have been accepted by many investigators.

Sonesson (1960) agreed with Wustrow on the course of the muscle fibers, as well as the correlation between their origin and insertion, but he claimed that the vocalis muscle could not be divided into two groups anatomically. Nakamura (1965), who reported the absence of Goerttler's portio thyreovocalis, claimed functional importance for the fibers running from the arytenoid cartilage to the vocal ligament. Quite unique is the finding of Zenker (1964), who claimed that most of the muscle fibers do not go all the way from the thyroid cartilage to the arytenoid cartilage. According to him, many fibers end within the muscle and are connected with each other there. There are various kinds of muscle-fiber connections. One of them is a chain of muscle fibers with intermediate tendons. Another shows insertion of some muscle fibers into the perimysium (a sheath of connective tissue) of one or of several muscle fibers. Thus, fibers of the vocalis muscle form complicated networks which are quite different from the parallel arrangement of the fibers as described by Wustrow or by Sonesson. Zenker concluded that the complex networks of muscle fibers contribute greater plasticity to the vocal folds.

In the vocalis muscle, many investigators agree on the existence of another group of muscle fibers rising from the arytenoid cartilage and inserting into the conus elasticus, which is a membranous, elastic connective tissue extending from the under surface of the vocal folds to the subglottal wall. Takase (1964), in his comparative anatomical study, claimed that this muscle-fiber group was phylogenetically more important than the portio aryvocalis. Schlosshauer and Vosteen (1957) and Hiroto (1966) have pointed out that contraction of these muscle fibers contributes to changing the shape of subglottal wall, which is shown in a frontal X-ray tomogram. Tomograms show the subglottal wall to be dome shaped during phonation but much less so during respiration.

Functional characteristics of the individual laryngeal muscles have been studied by examining contraction properties of the muscles. This can be done by recording the mechanical force or displacement caused by the contraction of the muscle. A single twitch contraction takes place as a response to a single stimulation applied either directly to the muscle or to the muscle nerve. When both of the muscle endings are fixed, thus keeping the muscle length constant, muscle tension is increased by an isometric contraction of the muscle. When one end of the muscle is pulled by a constant external force, the muscle length is decreased by an isotonic contraction. We can record these mechanical changes along the time course of the muscle contraction. Time from the beginning to the peak of the twitch contraction is called contraction time, and from the peak to the end of the contraction, relaxation time. There are two kinds of muscles, fast muscles with shorter contraction time and slow muscles with longer contraction time. Relaxation time is shorter in fast muscles and longer in slow muscles. Repetitive stimulation evokes a chain of overlapping twitches which result in a greater contraction. With an increase of the frequency of repetitive stimulation, more overlapping takes place between twitches and finally a complete mechanical fusion of the individual twitches (complete tetanus) is reached. Minimum frequency of stimulation to obtain complete tetanus of the muscle is called fusion frequency. Fusion frequency is higher for fast muscles and lower for slow muscles. The ratio between maximum tetanus tension and maximum twitch tension (in isometric contraction) is called tetanus-twitch ratio. A greater tetanus-twitch ratio implies a wider range of variation in the extent of contraction of the muscle (Mårtensson, 1968). These measures represent main mechanical properties of muscles.

Mårtensson and Skoglund (1964) measured contraction time in isotonic contraction of the laryngeal muscles of the dog and other animals. In the dog, contraction time was 14 msec for the thyroarytenoid, 16 msec for the lateral cricoarytenoid, 30 msec for the cricothyroid, and 35 msec for the posterior cricoarytenoid muscle. Similar results were obtained for the thyroarytenoid and the cricothyroid muscles of the cat, the rabbit, and the monkey, while the contraction time was longer for muscles of the sheep. According to Mårtensson and Skoglund, contraction of the adductor muscles of the larynx, i.e., the thyroarytenoid and the lateral cricoarytenoid, is very fast and is surpassed only by the external eyeball muscles (8-10 msec). Among the laryngeal muscles, the cricothyroid and the posterior cricoarytenoid are relatively slow. They also measured tetanus-twitch ratio (Te/Tw) and fusion frequency (F.F.) for isometric contraction of the thyroarytenoid and cricothyroid muscles of the dog. The values are 400 pulses per sec (F.F.) and 10:1 (Te/Tw) in the thyroarytenoid muscle and 300 pulses per sec and 5:1 in the cricothyroid muscle. They claimed that fast contraction of the adductor muscles is important in closing the glottis fast enough for effective protection of the respiratory organ against foreign bodies and that the capacity for quick changes in muscle tension is also useful for the regulation of voicing.

Hast (1966a, 1967) also examined the mechanical properties (in isometric contraction) of the cricothyroid and thyroarytenoid muscles of the dog and the thyroarytenoid muscle of the cat. His data for the dog were similar to those of Mårtensson and Skoglund, i.e., contraction time of the thyroarytenoid muscle was very short (14 msec), while that of the cricothyroid muscle was longer (39 msec). On the other hand, contraction time of the thyroarytenoid muscle in the cat was 22 msec, which is longer than that of the dog and does not agree with the results of Mårtensson and Skoglund. Hast concluded that the thryoarytenoid muscle contributes to rapid adduction of the glottis as well as to regulation of vocal-fold tension and length, while the cricothyroid muscle serves primarily as a regulator of gross vocal-fold tension and length. He justified his finding of longer contraction time in the cat by arguing that phylogenetically the larynx of the cat is more primitive than that of the dog. In a later report, Hast (1969) examined the Hynan gibbon, the rhesus macaque, and the squirrel monkey in comparison with the dog and the cat for the contraction properties of the cricothyroid and thyroarytenoid muscles. Values obtained from these primates agreed with those of the dog but not of the cat. Muscles of the cat were slower than for the other animals examined, although a similar relationship was found between the cricothyroid and thyroarytenoid muscles. This pattern of species difference is consistent with differences in laryngeal structure. The larynx has a double sphincter (superior and inferior, i.e., false folds and true vocal folds) in the dog, the monkey, and man, while there is a single-valve sphincter in the cat. Hast then concluded that experimental results obtained from the dog's larynx, which is anatomically and physiologically closer to the larynx of primates, would be more valuable for the study of the human larynx.

Hirose et al. (1969) also reported on the contraction properties of the cricothyroid, thyroarytenoid, and posterior cricoarytenoid muscles of the cat. Data for the cricothyroid and the thyroarytenoid were similar to those of Hast. The cricothyroid muscle was slower than the thyroarytenoid in both contraction time and fusion frequency, but the posterior cricoarytenoid muscle was as fast as the thyroarytenoid.

The contraction properties presented by the several authors here are summarized in Table 1. Also in the table are measurements of extrinsic laryngeal muscles of the dog (Hast, 1968) and fast and slow muscles of the extremities. In general, intrinsic muscles can be classified as fast muscles; among them, the cricothyroid is slower than the others. The laryngeal muscles of the cat are slower than those of the other animals examined. The data seem to be consistent with the results of Dunker (1968) in whose experiment submaximal stimulation of the recurrent nerves of the dog caused a visible adduction of the vocal folds after each stimulus up to a frequency of 60 pulses per sec. At this frequency, however, the amplitude of oscillatory movement of the vocal fold decreased to one-tenth of the maximum excursion. For a higher frequency of stimulation the vocal folds remained still in the median plane of the larynx.

Biochemical and histochemical studies in combination with examinations of microstructures have revealed certain other aspects of basic muscle features. There are generally known to be two kinds of muscles (or muscle fibers), red and white. The red muscle contains a large amount of myoglobin and carries on an aerobic metabolism. A tonic, or sustained, contraction is a functional characteristic of this muscle. The white muscle, which shows phasic or transitional contraction, contains a small amount of myoglobin and carries on an anaerobic metabolism. Human muscles are mixtures of white and red muscle fibers. It is said that a predominantly white muscle is functionally more phasic and a predominantly red muscle, more tonic.

Gerebtzoff and Lepage (1960) concluded from their histochemical studies that intrinsic laryngeal muscles are not suitable for rapid response to stimuli but rather for tonic contraction under control of the proprioceptive reflex mechanism.

TABLE 1: MECHANICAL PROPERTIES OF THE LARYNGEAL MUSCLES

From Mårtensson and Skoglund (1964):

	C. T. (msec)	F. F. (spc)	Te/Tw
Cricothyroid			
Dog	35	>300	5:1
Cat	15-20	---	---
Monkey and Rabbit	25-30	---	---
Sheep	50-60	---	---
Thyroarytenoid			
Dog	14	>400	10:1
Cat	9-13	---	---
Monkey and Rabbit	8-10	---	---
Sheep	15-20	---	---
Lat. Cricoarytenoid			
Dog	16	---	---
Post. Cricoarytenoid			
Dog	30	---	---

From Hirose et al. (1969):

(Cat)	C. T. (msec)	H. R. T. (msec)	F. F. (spc)	Te/Tw
Cricotyroid	44	41	44	4.3:1
Thyroarytenoid	21	27	$\geqq 100$	7.5:1
Post. Cricoarytenoid	22	29	$\geqq 100$	3.4:1

Note:

C. T. (contraction time): Time from the beginning to the peak of the twitch tension.

H. R. T. (half relaxation time): Time for decay of the twitch tension from the peak to one half of the peak tension.

F. F. (fusion frequency): Minimum frequency of repetitive stimulation for complete tetanic contraction.

Te/Tw (tetanus twitch ratio): Ratio of the maximum tetanus tension to the maximum twitch tension.

Continued

Table 1 (cont.)

From Hast (1969):

	C. T. (msec)	H. R. T. (msec)	F. F. (spc)	Te/Tw
Cricothyroid				
Hynan Gibbon	39	36	43	3.4:1
Rhesus Macaque	36.4	23.2	49	4.4:1
Squirrel Monkey	18.8	13.6	83	4.6:1
Dog	39	32.3	45	4.2:1
Cat	52.8	42.4	37	3.7:1
Thyroarytenoid				
Hynan Gibbon	16	17	95	5.9:1
Rhesus Macaque	14	14.4	112	7.1:1
Squirrel Monkey	13.2	14	116	6.0:1
Dog	14	18	114	6.2:1
Cat	22	20	73	5.0:1

Contraction Time in Extrinsic Laryngeal and Other Muscles:

	C. T. (msec)	Author
Extrinsic Laryngeal Muscles		
Thyrohyoid (dog)	52	Hast
Sternohyoid (cat)	50	Hast
External Eye-Ball Muscles		
Rectus Lat. Oculi (dog)	8-10	Mårtensson and Skolund
Rectus Med. Oculi (cat)	7.5-10	Cooper and Eccles
Limb Muscles		
Tibialis Ant. (dog)	25-30	Mårtensson and Skolund
Tibialis Ant. (cat)	19-24	Gordon and Phillips
Crureus (cat)	58.2	Gordon and Phillips
Soleus (cat)	72.7	Gordon and Phillips

Studies on muscle microstructure have revealed that the mitochondrion, an intracellular granule, is relevant to cell metabolism; it is known that a large population of mitochondria indicates a higher rate of aerobic metabolism. Use of the electron microscope has made for great progress in the study of mitochondria. Berendes and Vogell (1960) found that the population of mitochondria is higher in the vocalis than in the cricothyroid and other laryngeal muscles. The authors concluded that the vocalis muscle receives a better energy supply and so can contract and relax more rapidly than the other muscles. Later, however, Berendes (1964) revised this opinion in agreement with the findings of Matzelt and Vosteen (1963) and Ganz (1962, 1964).

Matzelt and Vosteen (1963) have shown that patterns of enzyme activities indicate more aerobic metabolism in the laryngeal muscles of humans than in some skeletal muscles, such as the quadriceps and the sternocleidomastoideus. Among the laryngeal muscles, aerobic metabolism was highest in the posterior cricoarytenoid, while it was lowest in the vocalis muscle, and the arytenoid and cricothyroid were ranked between them. Two types of mitochondria were observed, larger ones on the surface of the muscle fiber and smaller ones between the myofibriles inside the muscle fiber. A dense population of mitochondria, especially the small type, was found in the posterior cricoarytenoid muscle. The other laryngeal muscles had far fewer mitochondria. Both metabolic and electron-microscopic studies showed a clear distinction between the posterior cricoarytenoid muscle and the other laryngeal muscles, while the vocalis muscle was not distinguishable from the others. Matzelt and Vosteen interpreted this as an indication of functional specialization of the posterior cricoarytenoid muscle, which is the only abductor muscle of the glottis working against the adductor muscle group. According to Bowden and Scheuer (1960), the muscle mass of the adductor group is four times as large as that of the abductor muscle. Matzelt and Vosteen pointed out that a certain part of the vocalis muscle showed a relatively dense population of mitochondria which might have given rise to a sampling error and, hence, an overestimation of the total number of mitochondria in the above mentioned work by Berendes and Vogell (1960).

Ganz (1962, 1964) studied the capillary blood supply and enzyme activity in the human laryngeal muscles. Capillary development was found to be much better in the intrinsic muscles than in the thyrohyoid muscle, which is one of the extrinsic muscles. Among the intrinsic muscles, the greatest capillary development was seen in the cricothyroid, followed by the arytenoid and the vocalis. In the vocalis muscle there was a fairly well-developed capillary supply for the part close to the vocal-fold margin. The location seems to be identical with the one where Matzelt and Vosteen, and probably Berendes and Vogell, found a relatively dense population of mitochondria. Ganz's enzymatic results were similar to those of Matzelt and Vosteen. Aerobic metabolism was most conspicuous in the posterior cricoarytenoid muscle and secondly in the arytenoid muscle, followed by the vocalis and the cricothyroid. Far less aerobic metabolism was found in the thyrohyoid muscle.

In their histochemical study of the larynx of the dog, Tomita et al. (1967) reported that the distribution ratio of red muscle fibers to white fibers is approximately 1:1 in every intrinsic laryngeal muscle, while it is 1:1.5 in the femoral muscle. They also found a lower rate of anaerobic metabolism in the laryngeal muscles than in the femoral muscle but higher than in the heart muscle or the diaphragm. They could not find significant differences among the laryngeal muscles. From the histochemical evidence, Tomita et al. concluded that the intrinsic laryngeal muscle is more tonic than the femoral muscle and more phasic than the heart muscle and the diaphragm.

Kawano (1968) also reported a biochemical and electron-microscopic study of the laryngeal muscles. In the dog, the laryngeal muscles consumed more oxygen than the femoral muscle and less than the heart muscle. Among the laryngeal muscles, oxygen consumption was greatest in the cricothyroid muscle, moderate in the posterior cricoarytenoid, and least in the vocalis muscle. The laryngeal muscles of the dog contained more mitochondria than the femoral muscle and fewer than the heart muscle. Among the human laryngeal muscles, oxygen consumption was greatest in the arytenoid muscle, followed by the vocalis, the posterior cricoarytenoid, the thyroarytenoid, and the lateral cricoarytenoid muscle in descending order. The number of mitochondria was highest in the arytenoid muscle, followed by the posterior cricoarytenoid, the vocalis, the cricothyroid, and the lateral cricoarytenoid in descending order.

In summary, these biochemical and electron-microscopic studies show that aerobic metabolism is more active in intrinsic laryngeal muscles than in skeletal muscles, and in this respect, the laryngeal muscles may be ranked between the skeletal muscles and the heart muscle. Among the laryngeal muscles, a higher aerobic metabolism in the posterior cricoarytenoid muscle has been reported by some investigators, while none of the reports claimed any special character for the vocalis muscle or the cricothyroid muscle. Functional characteristics of the laryngeal muscles implied by these data seem to be quite different from those indicated by the data on mechanical properties. At this point, we should note the discussion of Matzelt and Vosteen (1963), who claimed that one must not expect easy correspondence between metabolic findings and functional characteristics, such as white vs. red or phasic vs. tonic, for the laryngeal muscles. Mårtensson (196[illegible]) also warned that direct inference from the findings in microstructure, such as the mitochondrion, to the mechanical properties of the laryngeal muscles might be risky. This caution still seems to be warranted. Further research is needed to reach a deeper understanding of the nature of the laryngeal muscles.

Functional Anatomy of the Cricoarytenoid Joint

Movement of the two arytenoid cartilages is the only factor that is relevant to the opening (abduction) and closing (adduction) of the vocal folds. Each arytenoid is connected to the cricoid cartilage by way of the cricoarytenoid joint.

Studies on the functional anatomy of the cricoarytenoid joint in human larynges have been reported by Sonesson (1958), von Leden and Moore (1961), Frable (1961), and Takase (1964). All of these investigations have yielded similar results and point out a misconception in some textbooks which describe a rotatory movement of the arytenoid cartilage around the vertical axis to the joint surface.

According to the investigators just cited, the articular facet on the upper posterior surface of the cricoid cartilage forms a prominence which serves as the convex surface of a cylinder joint. The axis of the cylinder, which defines the longitudinal dimension of the joint, runs in the direction from dorsomediocranial to ventrolaterocaudal. The facets of the joints in the left and right cartilages are not always symmetrical; they often show some difference in size and in direction of the longitudinal axis. The arytenoid facet of the joint is located in the lateral part of the cartilage on the undersurface of the muscular process. This facet has a concave surface which is well adapted to the convex surface of the facet on the cricoid cartilage as shown in Figure 1. The transverse dimension of the arytenoid facet is larger than that of the cricoid, while its longitudinal dimension is smaller. Results of the measurements are summarized in Figure 2 and Table 2.

TABLE 2: FUNCTIONAL ANATOMY OF THE CRICOARYTENOID JOINT

	Sonesson	Frable and Moore (in Von Leden and Moore)	Takase
Cricoid Facet			
Longitudinal Dimension	5.8 mm	6.08 mm	6.7 mm
Transverse Dimension	3.8	3.48	4.2
Radius of the Surface	3.3	---	---
Arytenoid Facet			
Longitudinal Dimension	4.0 mm	3.46 mm	4.6 mm
Transverse Dimension	5.3	4.58	5.3
Radius of the Surface	3.7	---	---
Angle of the Joint Axis to			
Saggital Plane	24°	26°	---
Horizontal Plane	35	36.8	54°
Frontal Plane	41	40.6	23

The structure of the cricoarytenoid joint permits the arytenoid cartilage two principal types of motion. One of them, the main type, is a rotating motion around the longitudinal axis of the joint; the other is a longitudinal sliding motion parallel to the axis. Von Leden and Moore described another type of motion, though very limited in its extent, around the attachment of the posterior cricoarytenoid ligament to the cricoid lamina. Contrary to the descriptions in some classical textbooks, there is no rotating motion around the vertical axis. Because of the distance of the vocal process from the axis of the joint, the inward (adducting) and outward (abducting) rocking motions of the arytenoid around the longitudinal axis provide leverage to open and close the glottis. Adduction of the arytenoid cartilages not only closes the glottis but also lowers the level of the vocal process. The longitudinal

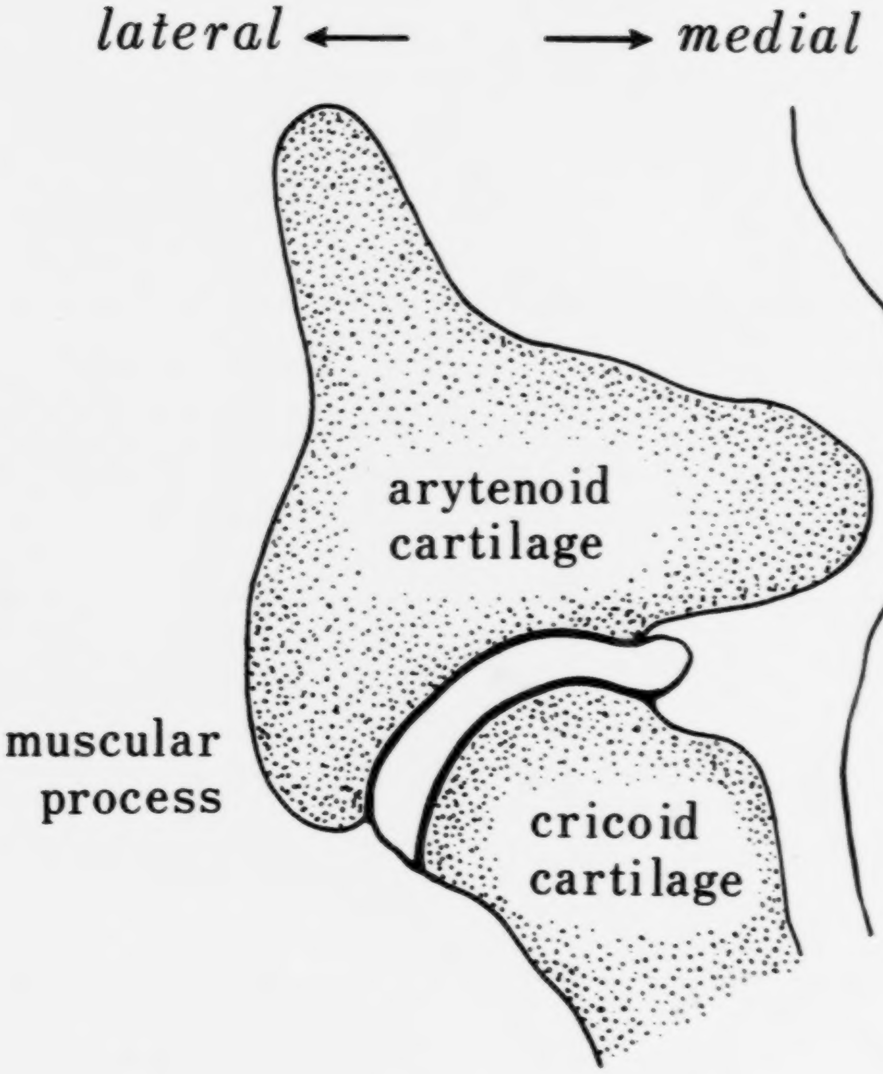

lateral ⟵
⟶ medial
arytenoid cartilage
muscular process
cricoid cartilage

FIGURE 1

Cross Section of the Cricoarytenoid Joint

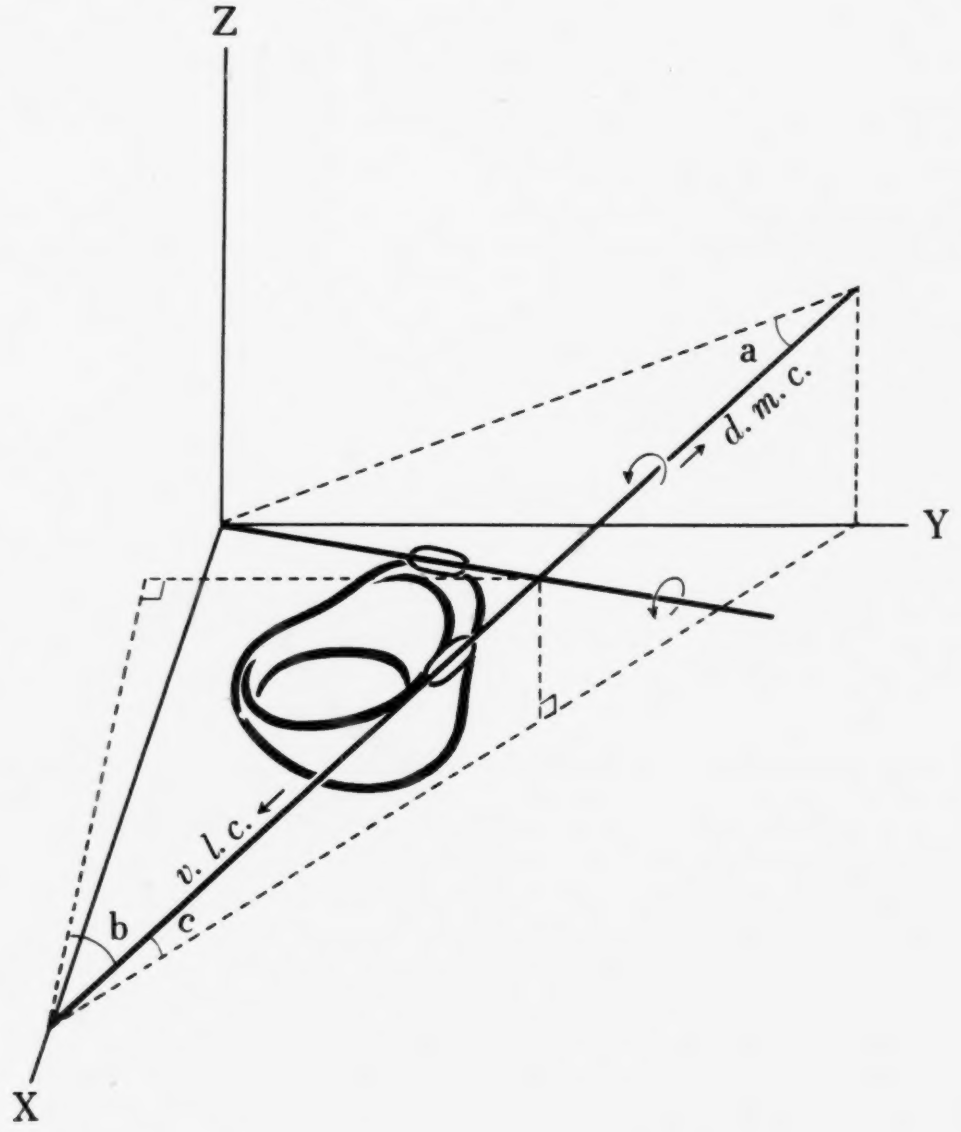


FIGURE 2

Axis of the Cricoarytenoid Joint

XY: horizontal plane; YZ: sagittal plane; XZ: frontal plane.
v.l.c.: ventrolaterocaudal direction; d.m.c.: dorsomedio-cranial direction.
a: angle of the axis to sagittal plane; b: angle of the axis to sagittal plane; c: angle of the axis to horizontal plane.

sliding motion of the arytenoid cartilages is very limited and of little importance for changing the glottal aperture. According to Sonesson, the maximum excursion of this movement is 2 mm, which is allowed by the difference in longitudinal dimension of the joint surfaces of the cricoid and arytenoid cartilages. This motion has an effect of shortening (relaxing) or lengthening (tensing) the vocal folds during vocal adjustment, with a slight lateral displacement.

Sonesson described the effects of individual laryngeal muscles on the movement of the arytenoid cartilages as follows: (1) the thyroarytenoid and the lateral cricoarytenoid muscles contribute to the adduction of the arytenoid cartilages and their linear sliding in the ventrolaterocaudal direction; (2) the arytenoid muscle causes the adduction of the arytenoid cartilages and their linear sliding in the dorsomediocranial direction; (3) the posterior cricoarytenoid muscle contributes to the abduction of the arytenoid cartilages with little effect on their linear sliding motion.

Reflex Control Mechanism of the Larynx

It is well known that auditory feedback is very important for the control of voice and speech production, which is a highly developed feature of human voluntary behavior. However, it is also true that in the process of voice and speech production, the activities of individual muscles of the larynx and upper articulatory organs are almost automatically, or subconsciously, adjusted and coordinated both with each other and with the respiratory muscles, to achieve the intended vocal or speech performance. This automatic adjustment of muscle activity is essentially similar to the regulation of coordinated movements in our voluntary actions (and those of other animals), such as walking, running, chewing, swallowing, and breathing. For limb muscles and other skeletal muscles that are innervated by the spinal nerves, the reflex control mechanisms of muscle activity have been fairly well elucidated by many investigators. The structures most important for this reflex mechanism are mechanoreceptors (specific endorgans of sensory nerves which respond to mechanical stimuli) in the muscle and its tendon. Afferent (sensory) nerve fibers from the receptors pass through the same nerve trunk as the efferent motor fibers innervating the muscle in which the receptors are located and are connected to the peripheral motor nerves in the spinal cord. Skeletal joints also have mechanoreceptors which contribute to the reflex control mechanism. Among these receptors, the most important and interesting one is the muscle spindle, a spindle-shaped structure in a capsule of connective tissue and lying in parallel with the muscle fibers. Inside the capsule there are two kinds of nerve endings, primary and secondary, and special muscle fibers which are called intrafusal muscle fibers. The primary ending belongs to a sensory cell with a large diameter (Group 1-a) nerve fiber directly connected to the peripheral motor cell, which innervates the muscle in which the same spindle is located. Thus, a circuit for monosynaptic reflex is formed. The secondary ending has an afferent fiber of small diameter (Group 11) which connects to the motor cell by several interneurons, thus forming a polysynaptic reflex circuit. The intrafusal muscle fibers are innervated by special motor cells with efferent fibers, which are smaller in diameter than those for the ordinary (extrafusal) muscle fibers. The muscle spindle responds to a passive extension of the muscle, and impulses from the primary ending facilitate the activity of the same muscle monosynaptically. The same afferent impulses inhibit the activity of the antagonist muscle via a polysynaptic reflex circuit. A popular example of this monosynaptic stretch reflex is the knee jerk evoked by tapping the tendon of the quadriceps femoris muscle at the knee joint. The response threshold of the secondary ending is higher than that of the primary, and the reflex pattern is more complicated.

The impulses from the muscle spindle are evoked not only by the passive stretch of the muscle but also by applying some chemical substances, such as succinylcholine, acetylcholine, or physostigmine. The response of the muscle spindle is inhibited by active contraction of the muscle.

The receptor located in the tendon also responds to mechanical stretch. The afferent fiber, the diameter of which is classified as Group 1-b, is connected to spinal motor cells via two or three interneurons. The tendon receptor responds to active contraction of the muscle as well as to passive stretch. The response evokes a polysynaptic reflex, the pattern of which is inhibitory to the muscle (as well as to a synergic one) and facilitatory to an antagonist muscle. This receptor has a higher threshold than the primary ending of the muscle spindle and is not influenced by chemical substances such as succinylcholine.

It is obvious that those proprioceptive reflex mechanisms, together with complex coordinating mechanisms at a higher level, play an important role in regulating the activity of muscles innervated by spinal nerves. On the other hand, the situation is quite different for the muscles innervated by cranial nerves. It is said that the muscles innervated by cranial nerves are somewhat different from those innervated by spinal nerves, although both of them are striated muscles. Actually, the innervation patterns are much more complicated and irregular in the case of cranial nerves, and some of the muscles involved are not skeletal or weight bearing muscles. A classic understanding is that certain muscles innervated by cranial nerves, such as the facial, lingual, external ocular, and laryngeal muscles, lack muscle spindles. This is becoming less acceptable since clear evidence of the existence of muscle spindles has been reported for the external eye muscles. However, the question as to whether or not muscle spindles exist in the intrinsic laryngeal muscles has not yet been answered.

In an anatomical study of the human larynx, which supported the findings of Goerttler (1950) and Paulsen (1958), König and von Leden (1961) reported the existence of muscle spindles in the intrinsic laryngeal muscles, especially a large population along the free margin of the thyroarytenoid muscle. Lucas Keen (1961) also reported that he found muscle spindles in all intrinsic muscles of the human larynx. Very recently Baken (1969), in his histological study of the human larynx, reported the existence of muscle spindles in all but the cricothyroid muscle. However, these morphological findings have not gained general acceptance among investigators. The argument has been focused on whether or not these receptors are identical with the muscle spindles found in the skeletal muscles. Rudolph (1961) found a certain kind of receptor with a spiral nerve ending in the human laryngeal muscles. However, its structure differed from the typical muscle spindle, and he suggested that it was a primitive proprioceptor rather than a muscle spindle. Gracheva (1963), and Nakamura (1965) also, reported some kind of nerve ending but claimed that the typical muscle spindle did not exist in the intrinsic laryngeal muscles of the human larynx. The absence of muscle spindles in the intrinsic laryngeal muscles of the dog was reported by Mårtensson (1964). Abo-El-Enein and Wyke (1966a) reported many spiral nerve endings in the laryngeal muscles of the cat that differed from the typical muscle spindle.

From these reports, one can reach the following conclusion: although the existence of the muscle spindle is still under discussion, most of the morphological studies agree as to the existence of some kind of proprioceptor in the intrinsic laryngeal muscles. More important than morphological evidence, however, is the neurophysiological study of the reflex control mechanism as a basis for explaining the function of the receptors.

Esslen and Schlosshauer (1960) reported a neurophysiological study with patients undergoing laryngectomy. A single rectangular wave was used as the electrical stimulation applied stimultaneously to the superior laryngeal nerve (which contains sensory fibers from the larynx) and the recurrent laryngeal nerve (which contains motor fibers to the intrinsic laryngeal muscles). Electromyographic data were recorded from the vocalis, lateral cricoarytenoid, posterior cricoarytenoid, and cricothyroid muscles. Following the stimulation of the nerves, two kinds of response were recorded for all muscles except the cricothyroid. One was the direct response (primary response) to the stimulation of the recurrent laryngeal nerve and the other (secondary response) was the response with a delay of 20 msec. The secondary response was inhibited by a stronger stimulation. This pattern is identical with that of the monosynaptic reflex of the limb muscle evoked via the spindle afferent fiber (Group 1-a). From these results, Esslen and Schlosshauer claimed the existence of a myotatic (stretch) reflex mechanism originating in the muscle spindles of the human larynx, the afferent impulses of which are mediated by the superior laryngeal nerve.

Bianconi and Molinari (1962) recorded afferent impulses from the recurrent laryngeal nerve of the cat. The recurrent nerve was cut, and the recording electrode was placed at the peripheral end of the severed nerve. This is a standard procedure for recording afferent nerve impulses without interference of efferent impulses conveyed within the same nerve trunk. A thread was tied to the arytenoid cartilage and a weight of 3 to 5 gr was attached to the other end to provide appropriate conditions for passive movements of the vocal folds. Slowly adapting discharges were elicited by a passive traction of the thyroarytenoid muscle. The discharge was suppressed during the contraction of the muscle evoked by direct stimulation of the muscle, while it was increased by the contraction of the posterior cricoarytenoid muscle. Similar discharge was also observed during passive stretch of the posterior cricoarytenoid muscle. Inhibition and acceleration of the discharge occurred upon contraction of the same muscle and of the thyroarytenoid muscle, respectively. Direct mechanical pressure applied to the muscle by use of a special probe evoked the same discharge, and the location of the source of the discharge, i.e., the receptor, was thereby determined. The receptors were located mainly in a region midway along the anterior two-thirds of the thyroarytenoid muscle and close to the arytenoid cartilage in the posterior cricoarytenoid muscle. They concluded that the experimental results were highly suggestive of muscle spindle activity.

Hirose (1961) also recorded afferent impulses from the recurrent laryngeal nerve of the cat. The response was increased by a passive extension of the vocal folds in the antero-posterior direction and inhibited by a pressure applied to the anterior surface of the larynx. No marked change was observed after intravenous administration of Decamethonium solution, which has an activating effect on the muscle spindle. He concluded that the afferent impulses did not come from muscle spindles but from a certain pressoreceptor of slowly adapting nature existing in the fascia or other connective tissues of the larynx.

Mårtensson (1963, 1964, 1967), using the dog as an experimental animal, recorded primary and secondary responses evoked by electrical stimulation of individual muscle nerves of the larynx. He examined the nature of the secondary response and concluded that this response was not a monosynaptic reflex via the spindle afferent nerve fiber but was a direct reflection of the discharge from the motor cell via the same motor nerve fiber. From the internal superior laryngeal nerve of the dog, he recorded afferent impulses which occurred in response to the contraction of individual muscles. There were two types of response; receptor excitation and receptor inhibition. Both of the responses showed very slow adaptation. Most receptors responded to the contraction of two or three different muscles and showed various conbinations of excitatory and inhibitory effects. The most common combination was receptor activiation on contraction of the posterior cricoarytenoid muscle and receptor inhibition for contraction of the thyroarytenoid muscle. Mårtensson observed some atypical responses, such as inhibitory responses to a submaximal contraction changing to excitatory ones for a maximal contraction of the same muscle. For tetanic contraction of the muscles, one and the same unit was reciprocally influenced upon simultaneous contraction of the two different muscles. For example, unit discharges evoked by tetanization of the thyroarytenoid muscle were interrupted by the additional contraction of the posterior cricoarytenoid muscle. The location of the source of the impulses was determined by applying mechanical stimulation to the muscles and other structures. As a rule, the receptors were found outside the muscles and near the lateral aryepiglottic fold and arytenoid cartilages. Mårtensson also observed electromyographically that the tonic component of the laryngeal muscle activity was increased by mechanical stretch of the muscle or by contraction of other muscles. This response was abolished by severance of the internal superior laryngeal nerve. He concluded that there was no evidence for a reflex control mechanism involving the muscle spindle, and the responses he observed probably originated in proprioceptors at the cricoarytenoid joint. The afferent impulses from the receptors were conducted by the internal superior laryngeal nerve. For receptors in the lower part of the larynx, such as the cricothyroid joint, Mårtensson suggested the recurrent laryngeal nerve as the afferent route. Criticizing the report of Bianconi and Molinari, which deduced the existence of muscle spindles in the larynx from their experiment on the cat (which has been mentioned earlier in this section), Mårtensson pointed out that the afferent impulses in their experiment had not been recorded from separate muscle nerves but from filaments in the main trunk of the recurrent nerve, and therefore there was the possibility that they had obtained afferent responses from joint proprioceptors instead of from muscle spindles.

Kircher and Wyke (1964a) found corpuscular nerve endings in the joint capsules of the laryngeal cartilages in the cat. These endings were morphologically identical with the rapidly adapting receptor in the skeletal joints of the same animal. The receptors were especially numerous in the case of the cricoarytenoid and cricothyroid joints. An articular nerve branch was also found to enter the cricothyroid joint from the recurrent laryngeal nerve. A neurophysiological study of the articulatory reflex mechanism was also reported by the same authors (Kirchner and Wyke, 1964a,b). A change of activity following the electrical stimulation of the articular nerve was observed electromyographically in the laryngeal muscles. The usual response to the stimulation of the nerve entering the cricothyroid joint was a burst of motor-unit discharges in the thyroarytenoid, lateral cricoarytenoid, and cricothyroid muscles, with coincident reduction in the activity of the posterior cricoarytenoid muscle. Similar responses were obtained by the stimulation of the nerve entering the cricoarytenoid joint. High intensity stimulation of the joint nerves evoked a more diffuse and prolonged reflex, with laryngeal spasm and change in respiratory rhythm. This response was assumed to be a pain reflex evoked by the stimulation of fibers which originated from another kind of ending.

The threshold for the production of the articular reflex was higher than that of the motor nerves but considerably lower than that of the mucosal branch of the laryngeal (sensory) nerve. The articular reflex response could be obtained only with light general anesthesia; with increasing anesthesia the response was abolished. Passive movements of the cricothyroid joint in caudal and anteromedial directions provoked a brief response to the muscles for the beginning and the end of movement. The response pattern of the muscles was identical with that evoked by the electrical stimulation. The response to the passive movement was abolished by local anesthesia of the joint capsule. The characteristics of the laryngeal articular reflexes demonstrated in this study were transient alterations in muscle activity provoked by discharges of rapidly adapting mechanoreceptors. The authors concluded that the reflexes were involved in the normal phasic coordination of activity in the laryngeal muscles during respiration and phonation and that these phasic mechanoreceptor reflexes were distinct from, and supplementary to, the tonic myotatic reflexes operated from receptors located within the laryngeal muscles.

Abo-El-Enein and Wyke (1966b), following their morphological study, demonstrated neurophysiologically the nature of the laryngeal myotatic reflex mechanism in the cat. Passive stretch was applied to the muscles by threads with attached weights, and the response of the muscles was examined electromyographically. Two types of response were demonstrated: facilitatory and inhibitory. Facilitatory response was evidenced by a sustained increase of motor-unit discharge in the stretched muscles. The response was initiated by the stretch exerted by a weight of 5 gr and increased with increments in the stretch tension. Beyond 15 gr of stretch tension, the facilitatory response disappeared and was replaced by an inhibitory reflex response. The facilitatory response seemed to be reduced by active contraction of the stretched muscle. With a stretch tension greater than 15 gr, the inhibitory reflex was manifested as a sustained reduction of motor-unit discharge in the stretched muscle. Unlike the facilitatory response, the inhibitory reflex was not aborted but rather was augmented by active contraction of the stretched muscle. This suggested that the mechanoreceptors responsible for the inhibitory effects were different from those evoking the facilitatory effects. The myotatic reflexes, both facilitatory and inhibitory, were observed only during the stage of light anesthesia. In the stage of moderate anesthesia both types of reflexes disappeared, while the monosynaptic reflex of the limb muscles remained intact. From the sensitivity to general anesthesia, Abo-El-Enein and Wyke concluded that the laryngeal myotatic reflex, as well as the laryngeal articular reflex, is a polysynaptic reflex system and is different from the monosynaptic reflex system of the skeletal muscles having muscle spindles.

Based on his morphological and neurophysiological study, Wyke (1967) pointed out three varieties of laryngeal phonatory reflex:

1) Those from mucosal mechanoreceptors, producing occlusive reflexes
2) Those from articular mechanoreceptors, producing phasic tuning reflexes
3) Those from myotatic mechanoreceptors, producing tonic tuning reflexes

The latter two, for which the morphological and neurological evidence has been reviewed above, are classified as proprioceptive reflex systems. The first of these is evoked by excitation of exteroceptive receptors and is well known as the mechanism protecting the air passage. Some neurophysiological studies have suggested the possibility of a contribution of these exteroceptive receptors to the laryngeal control mechanisms of respiration and phonation.

Sampson and Eyzaguirre (1964) recorded afferent impulses originating from touch receptors in the larynx of the cat. The impulses from receptors located rostral to (above) the vocal folds were observed in the superior laryngeal nerve, and those from receptors caudal to (below) the vocal folds were observed in the recurrent nerve. Touch receptors were sensitive to an air jet as well as to a light touch, and they also responded to vibratory stimulations up to 400 Hz. The responses were influenced by contraction of the laryngeal muscles.

Suzuki and Kirchner (1968, 1969) also observed afferent impulses from touch receptors in the external superior and the recurrent laryngeal nerves of the cat. The response was also evoked by a mechanical pressure and a small air jet. In the recurrent laryngeal nerve there were observed other impulses from slowly adapting mechanoreceptors located in the submucous tissue on the under surface of the vocal folds. This receptor responded to distention of the larynx and to vibratory pressure. For a sinusoidal vibratory stimulation, the frequency of the response impulses was exactly the same as the frequency of stimulation. These responses were abolished by surface anesthesia of the larynx. From these data Suzuki and Kirchner posited a reflex control of the larynx originating in the mechanoreceptors which respond to air flow through the glottis, to variations in subglottal air pressure, and to changes in the length and vibratory motion of the vocal folds.

Now we have seen evidence of reflexive laryngeal control activities, originating in proprioceptors in the muscles (although they may not be muscle spindles) and joints, and also in exteroceptors in the mucous membrane and submucous tissues of the larynx, as proposed in various experimental studies. Wyke (1967), in pointing out three varieties of laryngeal reflex mechanisms as mentioned before, assumed a reflex servomechanism, centered at the level of the brain stem, which organizes complex, coordinated reflexes driven by mechanoreceptors in the upper articulatory organs, the respiratory organs, and the larynx. He summarized the phonatory process as follows: at first the laryngeal muscles are voluntarily controlled, through the central motor pathway from the cerebral cortex, and preset to the glottal condition so as to produce the desired sounds. Secondly, when a subglottal air pressure is exerted and the air flow through the glottis has been set in motion, laryngeal articular and myotatic reflexes promptly operate to adjust the laryngeal posture, which might otherwise be deflected by the air pressure, so as to restore the preset postion. Finally, once the sound becomes audible, further adjustments, both voluntary and reflex, are made with the aid of the auditory monitoring system and the phonatory servomechanism.

Dunker (1968, 1969), recording neuronal responses in the bulbar sensory (soritarius) and motor (ambiguus) nuclei evoked by stimulations of the superior and the recurrent laryngeal nerves of the dog and the cat, has demonstrated evidence of complex inhibitory and facilitatory neural networks involved in the automatic control mechanism of the larynx at the level of the brain stem.

It is said that, for the skeletal muscles, the mechanoreceptors within the muscles, i.e., the muscle spindles and tendon receptors, do not contribute to the conscious perception of static position or movement of the limbs and vertebral column, although they serve as information sources for the reflex control of muscle activity. On the other hand, mechanoreceptors in the joint capsules are said to send information for perceiving static position and movement, as well as for the reflex control of muscle activity. Exteroceptive mechanoreceptors, such as touch receptors in the skin and the mucous membrane, apparently contribute to a conscious perception of the physical interaction of the surface of a part of the body with an object in the environment or with another part of the body. This information from the mechanoreceptors contributes, with the aid of visual monitoring, to the voluntary control of skeletal muscles. By combining these voluntary controls, based on conscious perception, with automatic reflex mechanisms at the level of the spinal cord as well as in the higher centers for the coordination of movements, one can develop, or learn, the highly complex and well-organized movements of daily life.

We can assume similar processes for laryngeal adjustments in voice and speech production, if we assume that the role of various mechanoreceptors in the larynx is similar to that of the receptors in the extremities and that the role of auditory monitoring is like that of the visual system. Human voice production, both in speech and in singing, is a highly intellectual and social type of behavior, and the laryngeal control for this is one of the voluntary learned muscle adjustments. It is needless to say that auditory perception is important for the learning and monitoring of speech and voice production. However, this does not necessarily imply a lesser importance of the reflex control of the muscles and sensory information from the receptors in the larynx as compared to those of the extremities. There seems to be no reason to omit these laryngeal feedback systems even for the voluntary (Wyke, 1967) presetting of the appropriate laryngeal posture at the onset of utterances.

The real question is how to obtain concrete neurophysiological evidence for describing the laryngeal feedback mechanism in human speech and voice production. We have seen much experimental evidence for the reflex control mechanisms of the larynx, although the results of animal experiments may not readily be interpreted as evidence for the human mechanism. We can also utilize the knowledge obtained from neurophysiological studies on skeletal muscles. Those studies have considerably increased our understanding of control mechanism of the larynx. However, it must be pointed out that these are limited to basic features of laryngeal adjustments and are still insufficient to be taken as directly relevant to the control mechanisms of voice and speech. Further steps are necessary to reach the level of voice and speech production. For these steps we should be more careful in utilizing the results of animal experiments, which show certain species differences even for the basic features of the larynx. Comparative physiological experiments will be useful to find a proper way of applying experimental results to human physiology. Furthermore, we must determine to what extent one can transfer the knowledge of the servomechanisms of the skeletal muscles, which are innervated by spinal motor cells, to those of the laryngeal muscles, whose motor cells are in the brain stem. We also should make clear how the phonatory control mechanism is related to those of the respiratory and protective (occlusive) reflexes, which are more primitive and phylogenetically more important functions of the larynx.

RECENT PROGRESS IN OBSERVATION OF LARYNGEAL ACTION DURING THE PRODUCTION OF VOICE AND SPEECH

Observation of physiological processes involved in voice and speech production contributes indispensable data to experimental phonetics. It is well known that the laryngeal mirror, invented by M. Garcia in 1854, has helped us greatly in understanding the essentials of voice physiology and pathology. Some of the current experimental methods for studying laryngeal mechanisms in speech will be discussed in this section.

Electromyography of the Larynx

Electromyography, a method of recording action potentials of muscles, is one of the most important and commonly used techniques in the study of physiology and pathology of the neuromuscular mechanism. Since the latter half of the 1950's, a number of electromyographic studies on the human larynx have been reported. Recently, electromyography of the larynx has been utilized not only for the study of sustained phonation but also for running speech. Increased interest in laryngeal mechanism has invited various improvements in electromyographic techniques; some of the technical aspects may be worth mentioning here.

One intriguing technical problem is the placement of the electrode to pick up the action potentials from the target muscle. In general, there are two types of electrode in use, surface and needle electrodes. The surface electrode is a metal plate which is placed on the skin immediately above the muscle. The electromyogram (EMG) in this case is the record of the electric potential difference between two surface electrodes appropriately placed. The needle electrode, being inserted into the muscle, can pick up the action potentials more accurately from a specific muscle than can the surface electrode. A common type of needle electrode is a concentric needle electrode, which is a hypodermic needle with insulated wire in it. The central wire, its metal surface appearing only at the tip near the needle, serves as the signal pick up and the needle surrounding it as the grounded shield. A bipolar needle electrode has two parallel wires within the grounded needle. The EMG signal obtained by this electrode is the potential difference between the closely positioned tips of the two conductors.

Electromyographic techniques developed by Faaborg-Andersen (1957) for studies of the human larynx provided a standard technique to subsequent investigators. In 1965 he summarized his techniques and his data together with those of others. He introduced single-wire concentric needle electrodes through the mouth to the vocalis, the arytenoid, and the posterior cricoarytenoid muscles. A laryngeal mirror was used for monitoring electrode placement. Because of difficulty in placing the electrode, any systematic study of the lateral cricoarytenoid muscle was not performed. Electrodes were inserted into the cricothyroid muscle and the extrinsic laryngeal muscles through the skin of the neck. A bipolar needle electrode was also used in special cases.

Although many valuable data have been obtained by Faaborg-Andersen and subsequent investigators this way, placing the needle electrodes by way of the mouth imposes several experimental constraints:

1) Surface anesthesia is necessary over a wide area from the mesopharynx to the larynx; this may disturb the subject's natural performance.
2) The wire connecting the electrode to the amplifier passes through the pharynx and mouth, interfering with articulatory movements.
3) The electrode tends to be dislodged by the opening and closing movements of the glottis or by the vibratory motion of the vocal folds. The electrode can be fixed in place by a holder; however, the procedure also interferes with vocal and articulatory performances of the subject.
4) Vigorous movements of the larynx, as in coughing and swallowing, must be avoided.
5) Under these conditions, EMG data can be obtained only over a short time span. In addition, it is impossible even to introduce the electrodes in some subjects who have a sensitive pharyngeal reflex.
6) Simultaneous recording from more than one of the muscles to be reached through the mouth is very difficult.

To overcome these difficulties, extralaryngeal approaches have been developed. Greiner et al. (1958) also using a needle electrode, penetrated the skin of the neck and the cricothyroid membrane and directed the electrode upward and laterally to reach the vocalis muscle from the under surface of the vocal fold. They anesthetized the subglottal mucous membrane by an intratracheal installation of 1 percent tetracaine solution. Hiroto et al. (1962) have described their extralaryngeal approach, which was based on their topographic anatomical study of the human larynx. According to them, the lateral cricoarytenoid and the posterior cricoarytenoid muscle can be reached by inserting the needle electrode lateral to the midline of the cricothyroid membrane and directing it posteriorly, slightly laterally, and upward. The vocalis muscle is reached by way of the subglottic space as described by Greiner et al. To reach the arytenoid muscle, the electrode is pushed backward in the subglottic space. Surface anesthesia of the mucous membrane is necessary for penetration of electrodes into the vocalis and arytenoid muscles. Hirano and Ohala (1969) used a similar approach but with hooked-wire electrodes; they described characteristic EMG patterns that identify individual laryngeal muscles.

If we are not seriously concerned with the precise location of the electrode in the vocalis muscle, it is practical to reach the thyroarytenoid muscle through the cricothyroid membrane without penetrating into the subglottic space and, consequently, without anesthetizing the larynx. Anesthesia of the larynx is unavoidable when the arytenoid muscle is the target. The most difficult case for the extralaryngeal approach is the posterior cricoarytenoid. The electrode can be placed only in a small part of the muscle near the insertion to the muscular process of the arytenoid cartilage. Hirano and Ohala suggested the use of a curved needle for this muscle. Zboril (1965) penetrated the cricoid cartilage from the subglottic space to reach the posterior cricoarytenoid muscle.

Among the extrinsic muscles, the sternohyoid muscle seems to be the easiest one to reach. Other extrinsic muscles caudal to the hyoid bone from which EMG recordings have been derived are the sternothyroid, the thyrohyoid, the thyropharyngeus, and the cricopharyngeus. The complicated structure and the small size of the laryngeal muscles require a well practiced hand and excellent anatomical orientation for the placement of an electrode in the target muscle. Audiovisual monitoring of the EMG signal by observing the electrical activity on an oscilloscope and by listening to the amplified signal through a loud speaker is indispensable during the procedure.

Another problem inherent in picking up electrical activity in a small muscle is contamination of the signal by field potentials, or cross talk, from nearby muscles. A concentric needle electrode, the most commonly used type, seems not to be secure from the disturbances of field potentials. Buchtal et al. (1957) reported that single motor-unit potentials with amplitudes of more than 50 microvolts originated only from a source region within 1 mm of the electrode. Dedo and Dunker (1966), however, illustrated that in the dog's larynx under certain conditions spike potentials greater that 50 microvolts could occur at least 5 mm away from the tip of the electrode. The use of a bipolar electrode, which has two signal-conducting wires, has been recommended to minimize the field potentials, and this was confirmed by Dedo and Hall (1969) in an experiment of the larynx of the dog.

There is still another kind of technical problem in the electromyographic study of the larynx. Insertion of stiff needle electrodes, most of them 0.45-0.65 mm in outside diameter, causes considerable discomfort to the subject and consequently interferes with natural speech utterances. Movements of the larynx tend to dislocate the electrode. In order to reduce the subject's discomfort and increase the reliability of the recorded data, Basmajian and Stecko (1962) made use of thin, wire electrodes. A pair of thin, insulated wires is inserted into a hypodermic needle which serves as a guide cannula. The tip of each wire is bent back to lie against the needle shaft for a short distance. The needle is inserted into the target muscle and then withdrawn leaving the wires in the muscle. The wires are kept in place by the hooks at their ends; they cause little discomfort to the subject because they are very thin and flexible. Basmajian and Dutta (1961) used the hooked-wire electrode for the palatal muscles, and Shipp et al.

(1968), for the inferior constrictor and the cricopharyngeus muscles in laryngectomized patients. Hirano and Ohala (1969) reported the use of hooked-wire electrodes for the intrinsic laryngeal muscles. According to them, the use of the hooked-wire electrode had the following advantages and disadvantages in laryngeal electromyography.

Advantages:

1) The electrode stays in place much better than does the needle electrode.
2) It causes little discomfort to the subject once it is in place.
3) Functioning as a bipolar electrode, it yields better location of the electrical activity picked up and greater freedom from ambient noise.

Disadvantages:

1) It is more difficult to implant accurately, since it is impossible to retract and reposition the electrode once it has been pushed in too far or in the wrong direction. More practice is required to handle the hooked-wire electrode than the needle electrode.
2) For thin muscles, a distance of 2-4 mm from the tip of the guide needle to that of the active wire may cause incorrect placement of the wire tip even though the needle tip is correctly located. To avoid this, the needle should be inserted into the muscle as parallel to the muscle fibers as possible.

In experimental phonetics, most electromyographic studies are not aimed at examining individual motor-unit potentials but rather at observing activity of the muscle as a whole. For this purpose, it is desirable to pick up as many motor-unit potentials as possible within the target muscle. This requires a compromise with the necessity of avoiding field potentials from nearby muscles. A large surface area at the active tips of the bipolar electrode is helpful. With each increment in muscle contraction, there is an increase in spike potentials from different motor units which overlap each other, resulting in an interference voltage. The classical method of obtaining a measure of muscle activity has been to examine the number of spike potentials or the amplitude of the interference voltage in a raw EMG record. A better electromyographic display can be obtained by rectifying and smoothing the EMG signal; this is called an integrated EMG trace. Examples of integrated laryngeal EMG have been presented by Faaborg-Andersen (1957, 1965) and Sawashima et al. (1958). Recent progress in data processing by means of a computer has provided an averaging technique for EMG samples. The principle is to record many samples of EMG for repeated tokens of the same type of vocalization. Then, they are rectified, smoothed, and averaged on a computer with a certain line-up point chosen for all the EMG samples. The final data can be either printed out or displayed on an oscilloscope. Averaged EMG levels provide a more detailed measure of the pattern of muscle activity. Averaged EMG levels of the cricothyroid and vocalis muscles as a function of vocal fundamental frequency and intensity have been presented by Sawashima et al. (1969).

Ultrasonic Observation of Vocal-Fold Movements

Ultrasonic techniques have been applied to medical diagnostic methods in recent years. The main purpose is to detect pathologic changes, such as a malignant tumor or abcess, inside the body. The chief advantages of this method are that the procedure is harmless to tissue, rapid in obtaining results, and involves no discomfort to the patient.

Ultrasonic waves, i.e., high-frequency sound waves far beyond the human auditory range, can pass through various kinds of media, including tissues of the body. Ultrasonic energy is reflected whenever the sound beam passes from one medium into another, the two having appreciably different acoustic impedances. The efficiency of reflection depends on the impedance ratio at the boundary between media and on the angle of wave incidence relative to the boundary surface. At frequencies higher than 1MHz, transmission of ultrasound into the air is negligible and almost all of the sound energy is reflected at a tissue-air interface.

Pulse-modulated ultrasonic waves are used to determine the distance between the transducer and the reflecting surface by measuring the time for the echo of the transmitted pulse to be observed. The transducer transmitting the sound usually serves also as the receiver. The echo intensity is displayed on a cathode-ray oscilloscope (synchroscope) as a vertical deflection from the horizontal axis, which represents time. The distance along the horizontal axis between the transmitted pulse and the echo represents the distance between the transducer and the medium boundary. This kind of display is schematized in Figure 3. The depth of the tissue to the internal surface in the body, such as the pharyngeal wall and vocal-fold edges, can be measured by this method. For a moving boundary, the position of the echo on the scope fluctuates along the x-axis according to changes in the distance between the transducer and the boundary. The mode of display can be changed so that the peak of the echo intensity appears as a bright spot on the x-axis. Using the vertical sweep of the oscilloscope for succeeding pulse transmissions, we can obtain a string of bright spots representing the movement of the boundary as a function of time along the y-axis. Asano (1968) displayed the vibratory motion of the vocal folds in this way (Fig. 4). The frequency of the ultrasound was 5MHz, and the pulse repetition rate was 5000 per second. A transducer 10mm in diameter was pressed against the skin of the neck on the wing of the thyroid cartilage so as to direct the sound beam perpendicular to the vocal-fold edge. Further adjustments in the placement of the transducer were made, monitoring the echo pattern on the scope. Figure 4 shows a schematic drawing of the display which Asano calls an ultrasonoglottogram. From this curve, one can measure the displacement and velocity of the vibrating vocal fold as time functions.

Miura (1969), using Asano's technique, placed two transducers on both sides of the neck and recorded echo patterns from both the left and right vocal folds. The two transducers were also used to pick up pulse beams passing through the glottis into the opposite sides of the larynx. Transmission of sound energy through the glottis takes place in this experimental condition only while the vocal fold edges are in contact with each other, thus providing an indication of the closed phase in each vibratory cycle. This is displayed as an intermittent line along the vertical axis of the scope simultaneously with the echo patterns of the left and right folds (Fig. 5). Miura pointed out that the waveform of the ultrasonoglottogram could be varied by changing the direction of the transmitted beam

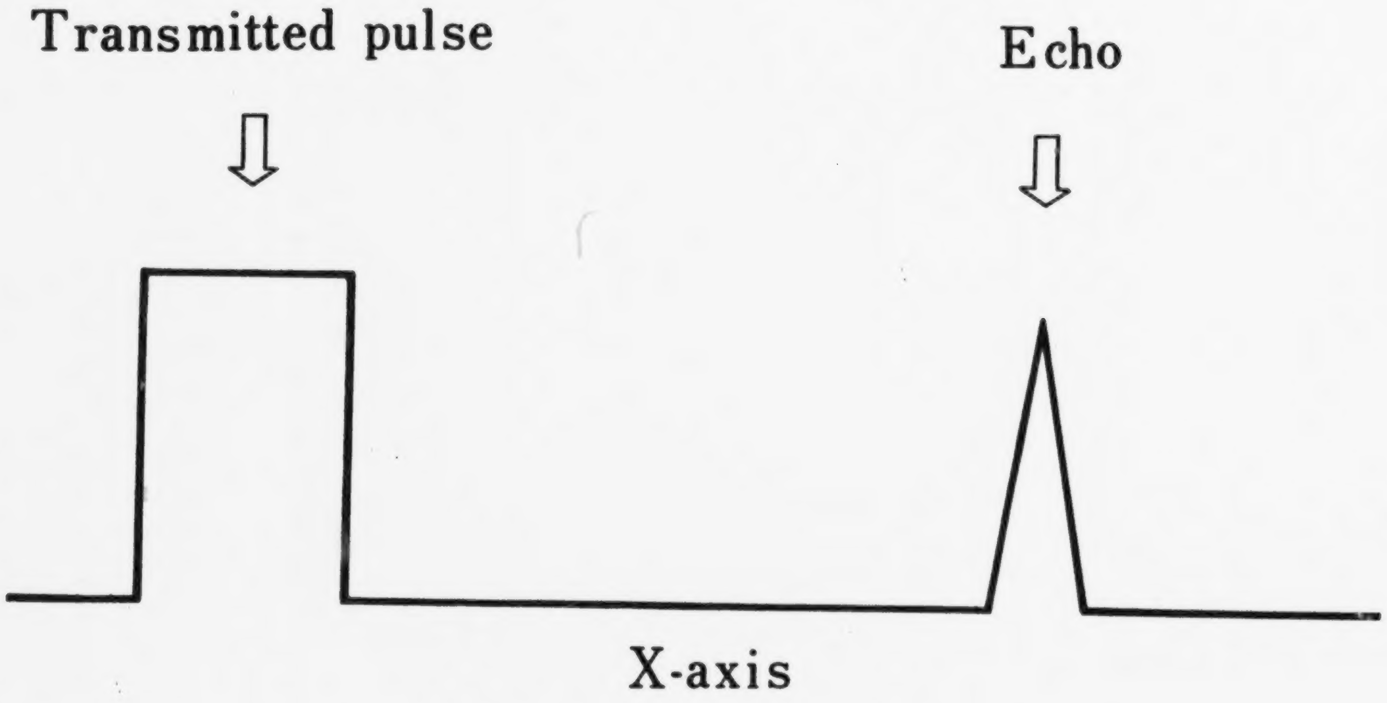

Transmitted pulse
Echo
X-axis

FIGURE 3

Display of Ultrasonic Echo

(See text for explanation)

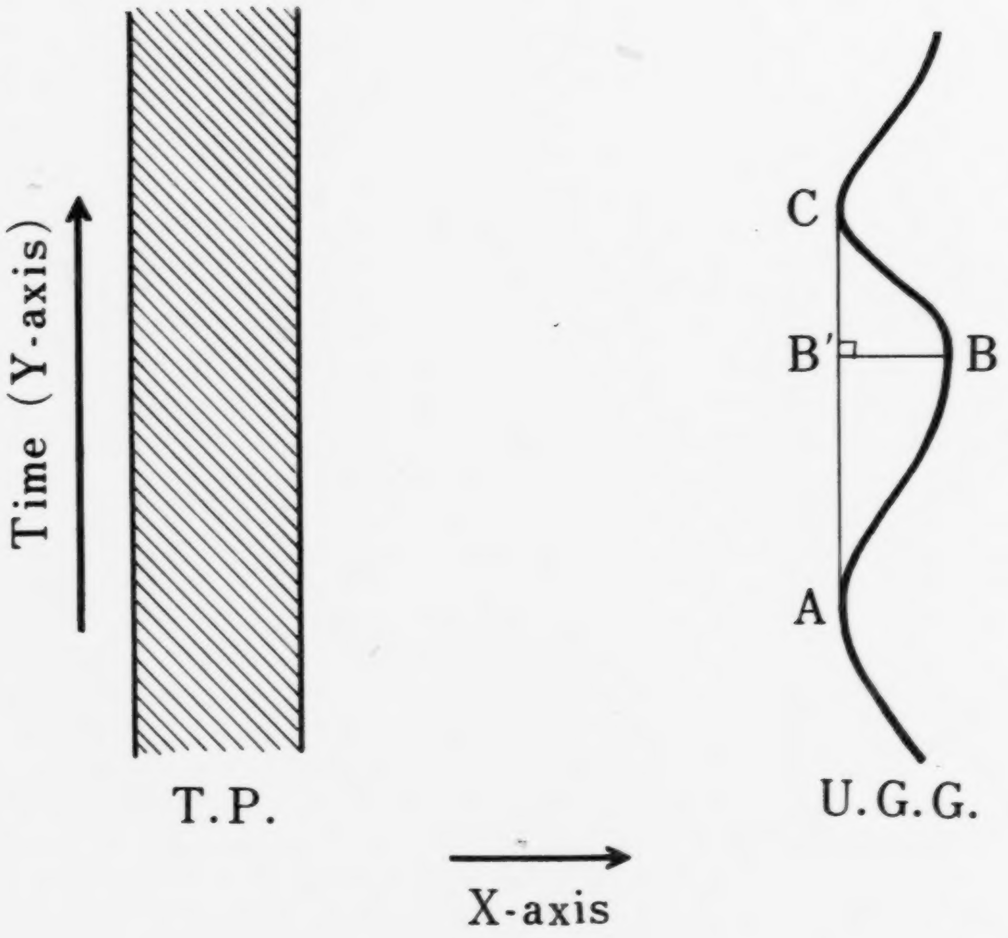

Time (Y-axis)
T.P.
C
B′
B
A
U.G.G.
X-axis

FIGURE 4

Ultrasonoglottogram (Asano, 1968)

T.P.: transmitted pulse; U.G.G.: ultrasonoglottogram; AC: one vibratory cycle of the vocal folds; AB: closing movement; BC: opening movement; BB': maximum amplitude of the vibration in horizontal plane.

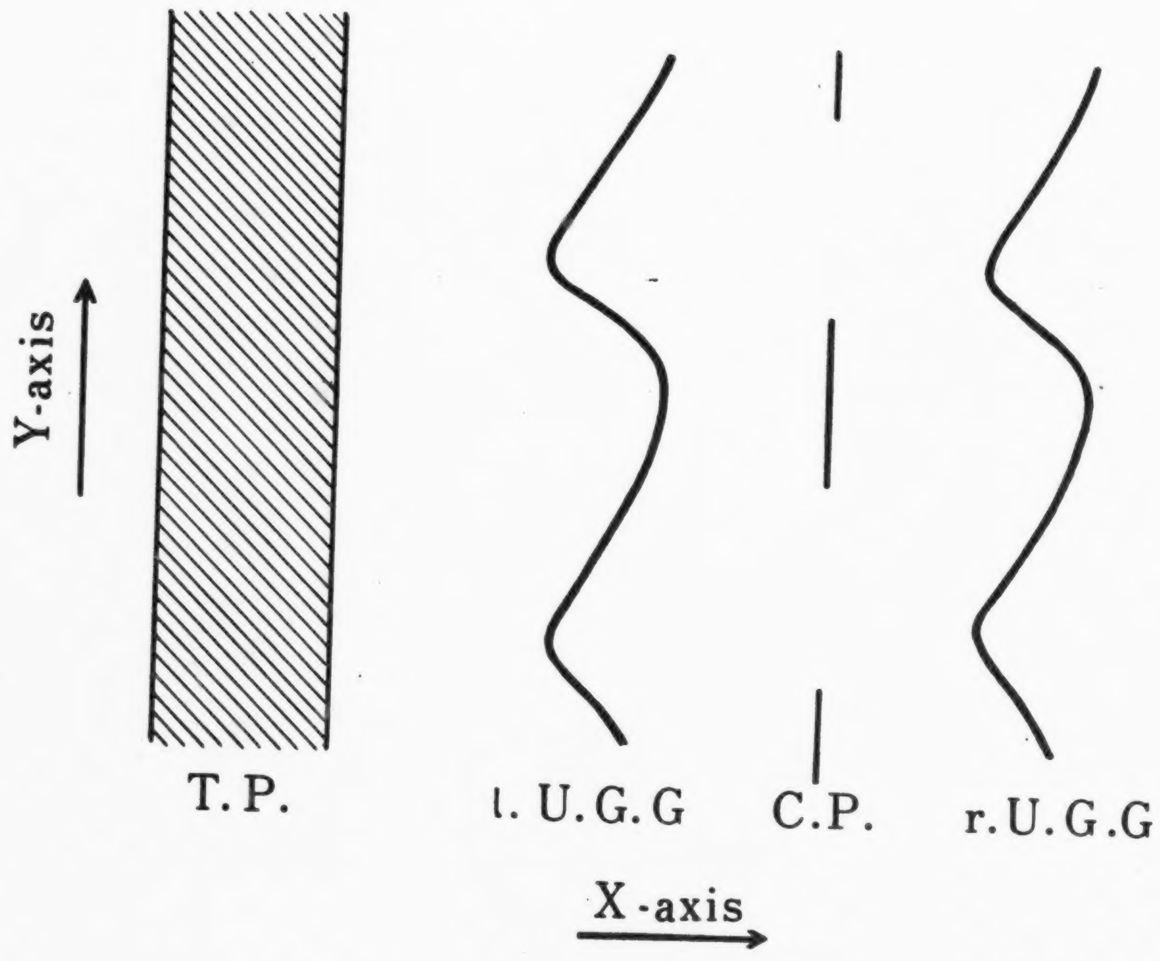


FIGURE 5

Simultaneous Display of Ultrasonoglottograms and Closed Phase During Vibratory Cycle (Miura, 1969)

T.P.: transmitted pulse; l. U.G.G.: ultrasonoglottogram of the left vocal fold; r. U.G.G.: ultrasonoglottogram of the right vocal fold; C.P.: closed phase.

slightly upward or downward. He also noted that the glottogram showed movements of the vocal-fold edges even during the closed phase of the cycle. He interpreted these results as an indication of a complex vibratory pattern of the vocal-fold edges which changed their shape during each cycle. He also explained that the echo recorded was not reflected from a fixed point but from a certain area of the vocal-fold edge. A technique similar to Asano's was applied by Kelsey et al. (1969) to detect lateral pharyngeal wall movements during speech articulation.

Ultrasonic Doppler velocity monitoring of the vocal-fold vibratory motion was reported by Minifie et al. (1968). This procedure is based on the frequency shift in the ultrasonic echo from a moving medium boundary. A continuous ultrasonic beam is directed into the neck, and the echo from the vocal-fold edge is recorded. When the vocal-fold edge approaches or moves away from the transducer, there is a shift in the frequency of the reflected sound. The difference in frequency between the transmitted and the reflected sound is called the difference frequency, or the Doppler frequency, and is approximately proportional to the velocity of the movement of the reflecting surface, in this case the vocal-fold edge. These investigators used a transducer with a diameter 8 mm, which transmitted ultrasound at a frequency of 5.3MHz. The relative displacement of the fold was calculated by integrating the velocity thus measured. They reported that the results of their preliminary experiments were in good agreement with those obtained by other methods, such as transillumination and high-speed motion pitures. Later, however, Beach and Kelsey (1969) pointed out discrepancies between the results of this method and data from high-speed motion pictures taken simultaneously. Discrepancies were found in the timing, magnitude, and general shape of the velocity and displacement curves; furthermore, the relationship between Doppler frequency patterns and the corresponding vocal-fold motions in the film was inconsistent and ambiguous. They concluded that the present ultrasonic Doppler technique does not provide a reliable measure of vocal-fold vibration and that the unreliability of this technique seems to stem from its lack of precision in analyzing complex and detailed motions of the vocal folds during phonation.

By way of summary, the following points about the ultrasonic techniques for observing vocal-fold movements may be made:

1) Since no instrument need be inserted into the vocal tract, the procedure causes no discomfort or disturbance to the subject.
2) The ultrasonic energy for this purpose is harmless to the subject.
3) Absolute values of the velocity and the displacement of the vocal fold are obtainable.
4) The major problem is that the correct interpretation of the obtained wave forms as the vocal-fold vibration patterns suffers from some difficulty. A complex local shape of the vocal-fold surface as the moving reflection boundary causes this difficulty. A more detailed examination of the results obtained by this method, perhaps by use of smaller transducers, in comparison with those obtained by other methods, such as high-speed motion pictures, is in order.
5) Another source of difficulty is the up-and-down movement of the larynx during utterances. No solution has been offered in this respect.

Transillumination of the Glottis

A technique for recording glottal-area variation by measuring the amount of light passing through the glottis was developed by Weiss (1914). In experiments by him and some followers, a light source was placed in the subglottic space of an excised larynx and the light passing through the glottis was recorded on photographic paper on a rotating drum. Rhomboidal marks were obtained on photographic paper which corresponded to the individual vibratory cycles of the vocal folds of the excised larynx.

Sonesson (1960) recorded the light passing through the glottis by a photo-electric device applied to the normal human subject as follows. A laryngo-diaphanoscope lamp, which is a modified microscope-illuminating DC light, is placed against the pretracheal wall of the neck in the fossa jugularis. A light-conducting rod is inserted through the mouth into the hypopharynx so that the tip of the rod is placed posteriorly and just below the tip of the epiglottis. Surface anesthesia is necessary for the placement of the rod. A photomultiplier tube is connected to the other end of the rod. The light from the lamp illuminates the subglottal space through the tissues of the neck. The illuminating light passing through the glottal aperture is transmitted through the light-conducting rod to the photomultiplier tube. The output of the photomultiplier tube is displayed on a cathode-ray oscilloscope as a function of time. The record is called a photo-electric glottogram and represents the area of aperture in the horizontal plane of the glottis as a function of time. Using this technique, Sonesson measured the open period, the opening phase, and the closing phase of the glottal vibratory cycle for sustained phonations in different pitch and intensity conditions. According to him, the results obtained from the glottograms were in good agreement with the results obtained from high-speed motion-picture films.

Sonesson's technique, introducing a stiff rod into the hypopharynx through the mouth, imposed severe limitations to the articulatory movements of the supraglottal organs. To avoid these limitations, Malécot and Peebles (1965), Ohala (1966), and Frøkjaer-Jensen (1967) have introduced modifications of Sonesson's technique. For developing the modified techniques, miniaturized solid-state photosensors have been of great help. The principle of the new techniques is to introduce a small photosensor attached to the tip of a thin flexible plastic tube through the nose into the pharynx, thus making transillumination possible during speech articulation. Slis and Damsté (1967) introduced a flexible bundle of glass fibers via the nose to transmit the light to the photosensor placed outside the subject.

Lisker et al. (1969) reversed the positions of light source and photosensor relative to the glottis. They introduced a small lamp into the hypopharynx via the nose and placed a photomultiplier tube on the pretracheal wall at the neck. Similar positions for illuminating the glottis and picking up the transglottal light have been used by Sawashima (1968). In his system, a flexible fiber-optics was inserted through the nose so that the transillumination record could be obtained simultaneously with a motion picture of the larynx during speech articulation. The illuminating light for the laryngeal photography also served for the transillumination. Sawashima reported a linear relationship between the output level of the phototube and the glottal area in the motion-picture films during the gross opening (abducting) and closing (adducting) movements of the vocal folds. He also noted that a shift of the output level could be caused by a change in the positioning of the optical cable, i.e., the illuminating light, relative to the larynx and, hence, emphasized the importance of visual or photographic monitoring of the glottis for the correct interpretation of the transillumination record.

These modifications of Sonesson's method have extended the application of the transillumination technique to studies on glottal adjustments, the abduction and adduction of the vocal folds as well as the cessation and resumption of glottal vibration during consonant articulations. It has been assumed that data obtained by the transillumination technique provide a good approximation of the glottal-area function, although it is impossible to calibrate the instrument to measure the absolute area of the glottis. Colman and Wendahl (1968), however, have raised some questions regarding the validity of this technique. They recorded the glottograms simultaneously with high-speed motion pictures during sustained phonations. Comparing the glottograms with the glottal areas obtained from the motion pictures for the same vibratory cycles, the authors found a considerable dissimilarity between the glottal wave forms obtained by the two methods. As possible error sources in the transillumination method, the following factors were pointed out:

1) The light-density distribution within the vocal folds may not be constant.

2) The changing cross-sectional area of the folds in an anterior-posterior plane may result in an uneven illumination of the folds.

3) Light reflections from the mucosal surfaces may not be invariant.

4) Vertical vocal-fold movements toward and away from the light source are not taken into account.

5) The location of the monitoring device causes different wave forms.

Thus, Colman and Wendahl concluded that, without acceptable validation of the procedure, the conclusions drawn from data obtained by transillumination alone must be regarded with skepticism, particularly regarding the spectral properties of the glottal wave.

It is known that the illuminating light passes not only through the glottal aperture but also through the laryngeal tissues to reach the photosensor. Thus, some conditions of the larynx other than the glottal aperture, such as variations in the thickness of the vocal folds, may also be reflected on the transillumination traces. On the other hand, it is also true that most of transillumination traces show wave patterns on which we can clearly observe the opening phase, the closing phase, and the closed phase of the glottal vibration. Examples of the traces for different types of phonation are shown in Figure 6. It seems, therefore, that we can attach considerable value to studies made using this technique, providing care is taken in its use.

Summarizing the above observations, the following points can be made about the transillumination technique:

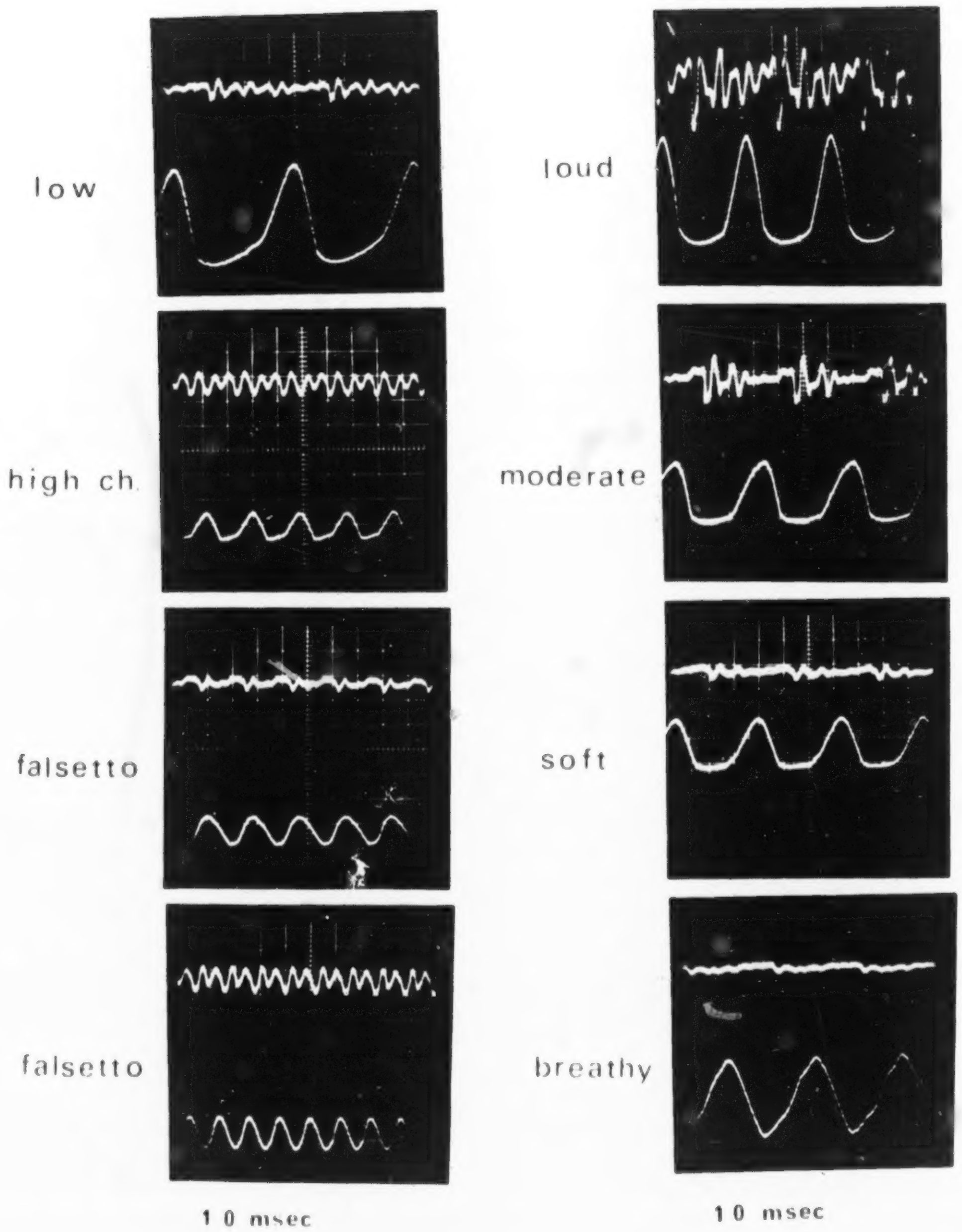


FIGURE 6

Simultaneous Recording of Transillumination (Lower Traces) and Voice (Upper Traces) for Phonation of /a/ in Different Pitch and Intensity Conditions

left column: variation in vocal pitch in chest voice (upper two records) and falsetto (lower two records).
right column: variation in vocal intensity (upper three records) and breathy phonation (lowermost record).

1) The technique has been developed as a straightforward method for recording glottal-area variations, in both vibratory cycles and gross abducting and adducting movements of the vocal-folds, during speech utterances.

2) The absolute area of the glottis is not obtainable by this technique.

3) Several sources of artefacts in the technique should be taken into account for the correct interpretation of the recorded data. Some of the artefacts may be caused by conditions in the larynx, such as variations in the thickness of the vocal folds. A shift in the positioning of the instruments relative to the larynx is also a major source of experimental artefacts. Conditions, such as up-and-down movements of the larynx, displacement of the instruments in the pharynx, and interruption of the light by the epiglottis during utterances should be carefully monitored to minimize errors.

Electrical Glottography

The technique of electrical glottography for registering glottal vibratory movements by measuring changes in electrical resistance across the neck was first reported by Fabre (1957). In this technique, a pair of plate electrodes were placed on the skin on both sides of the neck above the thyroid cartilage. A weak, high-frequency voltage of 0.2 MHz was applied to the electrodes, and a small fraction of the electric current passed through the larynx. The transverse electrical resistance of the larynx varied depending on the opening and closing of the glottis, and a modification in the amplitude of the transglottal current occurred in correspondence with the vibratory cycles of the vocal folds. The amplitude modification of the current was detected from which the electrical glottograms were produced. By means of this technique, Fabre (1958) examined glottal vibrations in different types of phonation. He further applied this technique to the observation of the opening (abduction) and closing (adduction) movements of the vocal folds in respiration (Fabre, 1961). Michel and Raskin (1969) have developed an improved version of the electrical glottograph, the electroglottometer Mark IV.

In electrical glottography, two plate electrodes are placed outside the neck, and no instrument needs to be inserted into the mouth or the pharynx, thus minimizing the discomfort to the subject. On the other hand, there seem to be several problems inherent to this technique. Among them, the following points were noted by van den Berg (1962):

1) The signal is too small to be detected in a subject with a heavy neck.

2) It is difficult to avoid contamination of the signal by microphone effects arising from variations in the contact resistance between the electrode and the skin due to vibration of the larynx as a whole.

Application of electrical glottography to voice and speech research have been attempted by several investigators. Fischer-Jørgensen et al. (1966) noted that some glottograms reflect possible degrees of glottal opening during speech articulations, while some do not. As possible artefacts affecting glottographic measurements, the authors pointed out variations in contact resistance between the electrodes and the skin, up-and-down movements of the larynx relative to the electrodes, and the variations in the conditions of the pharyngeal wall. In making comparisons between electrical

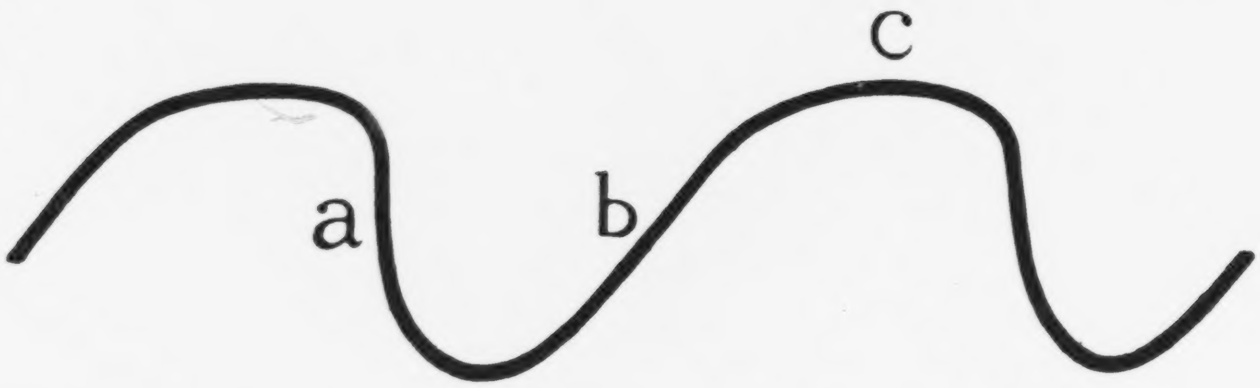


FIGURE 7

Wave Form of Electrical Glottogram

(See text for explanation)

and photo-electric glottograms, Frøkjaer-Jensen (1968) concluded that the opening of the glottis seemed to be better represented in photo-electric glottograms, whereas the closure of the glottis (in particular its vertical contact area) was probably better reflected in electrical glottograms. Fant et al. (1966) compared wave forms of electrical glottograms, photo-electric glottograms, and the volume-velocity source signal derived from the speech signal by an electronic inverse filter.

A schematic drawing of a typical waveform of an electrical glottogram for chest voice is shown in Figure 7, in which an ascending curve represents an increase in the electrical resistance and a descending curve a decrease. In the figure, a glottal cycle consists of an abrupt falling curve (a), followed by a gradual rise (b) and then a high plateau (c). In his original work, Fabre (1957, 1958) correlated the downward deflection of the curve to the maximum opening of the glottis and the plateau to the closed phase. This should be reversed as has been pointed out by Ohala (1966) and Fant et al. (1966). The comparative studies of the different types of glottograms have revealed the true interpretation of the curves. In Figure 7, the abrupt fall reflects the contact of the closing vocal folds with a rapid increase in the contact area in the horizontal and vertical directions. The gradual ascent reflects a rather slowly decreasing contact area after the maximum closure of the vocal folds. When the vocal folds have opened in their full length, no further separation appreciably contributes to the variation of the electrical resistance, as is observed in the plateau of the glottogram.

At present, the following points may be made in summarizing the reports on electrical glottography:

1) The procedure is carried out with a minimum of discomfort for the subject.

2) The record reflects the glottal condition during closure better than during the open period. The presence or absence of glottal vibration, as well as the fundamental frequency, can be readily determined.

3) The glottogram seems to be considerably affected by artefacts such as variations in the contact resistance between the electrode and the skin, up-and-down movements of the larynx relative to the electrodes, and conditions of the pharyngeal wall and other structures in the neck. It is difficult to estimate to what extent the glottal condition contributes to the electric impedance variations between the electrodes, and a quantitative interpretation of electrical glottograms seems to be less direct than, for example, photo-electric glottograms.

Fiberoptics for Viewing the Larynx

Although there are various methods for indirectly observing laryngeal action during speech articulations, visual examination is essential at least in establishing the correct interpretation of the data obtained by indirect methods. A laryngoscopic technique by use of a fiberoptics cable introduced into the pharynx via the nose will be briefly described here. The technique has been developed by Sawashima and Hirose (1968).

The fiberoptics, the outside diameter of which is approximately 6 mm, contains a group of glass fibers for transmitting images and another group

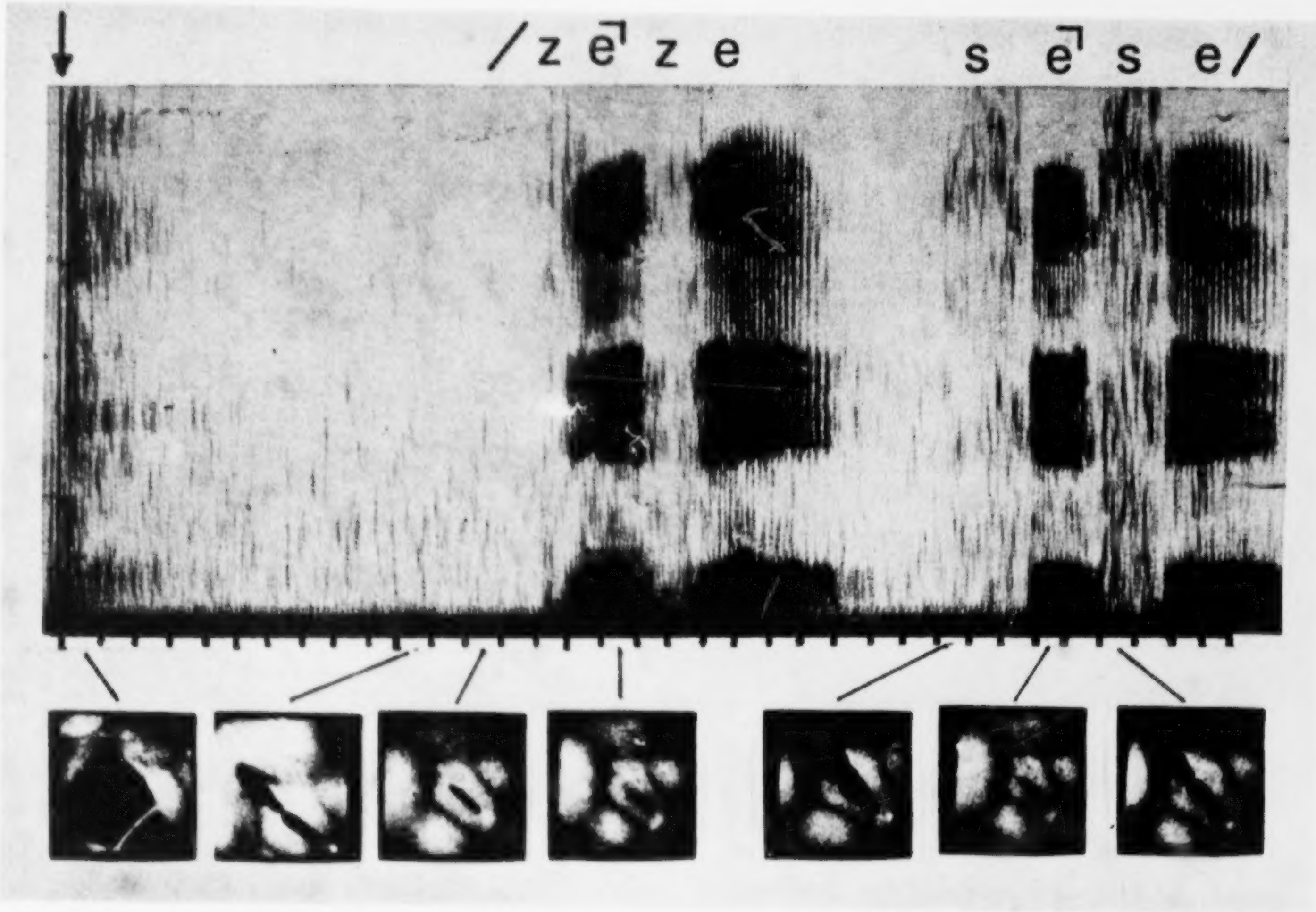


FIGURE 8

Sound Spectrogram and Selected Frames of the Motion Picture of the Larynx in a Pronunciation of /ze⌝ ze se⌝ se/ (Sawashima, 1968)

Short vertical bars under the spectrogram demarcate time intervals for successive frames, and the arrow at the upper left corner indicates the time mark for synchronization of sound and frame.

for conducting illuminating light. For introducing the optical cable through the nose, surface anesthesia on the mucous membrane of the nose and the epipharynx is necessary. The tip of the optical cable is placed in the hypopharynx and posteriorly to the tip of the epiglottis. The insertion of the optics does not cause any recognizable distortion in the natural speech of the subject. A 16 mm cine camera is attached to the eye piece of the optics, and simultaneous recording of the speech signal is made with synchronization time marks placed on both the film and the audio-tape. The use of a DC light source for illumination and a photosensor placed on the pretracheal wall makes a combination of the photo-electric glottography with the filming possible. Figure 8 shows selected frames of a film taken during articulations of CVCV syllables. The frames are displayed in correspondence with the sound specrogram. The frame rate of the film is 24 per second. In the selected frames, from left to right, the vocal folds are: in abducted (inspiratory) position before the utterance (first frame), in the course of adduction (second frame), in adducted (phonatory) position before voice onset (third frame), in phonatory position and actually in vibration for the vowel [e] of the first syllable of /zèze/ (fourth frame), set apart for the initial [s] of /sèse/ (fifth frame), in vibration for the following vowel [e] (sixth frame), and again set apart for the second [s] of /sèse/ (last frame). The blurred edges of the vocal folds in the fourth and sixth frames, in contrast to the others, evidence the vocal-fold vibration. By this technique, opening and closing movements of the vocal folds and the arytenoid cartilages during respiration and speech articulation are visually examined. The presence or absence of vocal-fold vibration is also detectable.

At present, there are several limitations in this technique:

1) The small diameter of the light guide limits the amount of illuminating light and consequently the frame rate of the motion picture. Therefore, a high-speed motion picture for analyzing details of vibratory movements of the vocal folds is not feasible with the present system.

2) Changes in the position of the tip of the fiberoptics relative to the larynx, which take place during some articulatory movements, may cause variations in the image size and the viewing angle of the larynx. Backward tilting of the epiglottis also interferes with the glottal view. In general, difficult sounds for obtaining good views are nasals and low back vowels.

3) There is no reference available for calibrating the absolute dimensions of the glottal aperture.

LARYNGEAL ADJUSTMENTS IN VOICE AND SPEECH PRODUCTION

In this section, some results of the observation of laryngeal action on voice and speech will be outlined.

Laryngeal Control Mechanism in Vocal Pitch and Intensity During Sustained Phonation

Numerous studies have been reported on laryngeal adjustments for vocal-pitch control. Experimental evidence supports the general concept that vocal pitch changes as a function of longitudinal tension in the vocal folds, the mass of the vocal-fold tissue involved in the vibration, and subglottic pressure. Vocal-fold tension increases with the contraction of the vocalis muscle and/or the extension of the vocal folds by external force. It is known that the laryngeal adjustments in the lower pitch range, which is called chest register, are different from those in the higher pitch range, the falsetto register. The range of conversational vocal pitch is normally included in the chest register.

In an experiment on excised human larynges, van den Berg and Tan (1959) observed an increase in fundamental frequency and a shift from the chest register to falsetto as extension of the vocal folds was increased by external force applied to the thyroid cartilage. They also noted that the extensibility of the vocal folds was almost entirely determined by that of the vocal ligaments and that the greatest part of the longitudinal force applied to the thyroid cartilage was taken up by the vocal ligaments. At large elongation of the vocal folds, tension in the vocal ligaments was much greater than that obtainable by the contraction of the vocalis muscle. Based on these experimental results, van den Berg (1960, 1962) concluded that, in the chest register, tension in the vocal ligaments was very slight and vocal pitch was determined by the contraction of the vocalis muscle associated with medial compression of the vocal folds, whereas tension in the vocal ligaments caused by the external force became dominant in falsetto.

Differences between the chest and falsetto registers, as observed in the vibratory motions of the vocal folds of living human subjects, have been reported by several investigators. According to Rubin and Hirt (1960), who made a high-speed cinematographic study, for example, the vocal folds in chest register vibrate along their full length, striking each other with their entire mass in vibration, whereas in falsetto, the vocal-fold margins become very thin and touch each other lightly or not at all, the main mass of the vocalis muscle remaining uninvolved in the vibration.

Hast (1966b), with electrical stimulation of the recurrent nerve of the dog, measured the tension of the thyroarytenoid muscle in its isometric contraction. The tension was observed to increase with an increase in the initial length of the muscle prior to stimulation, as well as with increases in the voltage and frequency of the stimuli.

Lateral radiographic studies on vocal-fold length in the singing of ascending scales were reported by Sonninen (1956, 1968), and maximum elongation was estimated at about 4 mm. The elongation curve showed a steep rise up to a certain limit and then gradually became saturated near the register shift from chest to falsetto and thereafter. Rubin and Hirt (1960) also studied lateral X-rays and noted considerable elongation of the glottal chink in passing from low to high chest voice, while there was no further lengthening in transitions to and within falsetto. Similar results were obtained by Damsté et al. (1968). Arnold (1961) reported elongation of the vocal folds by 5 mm, or about 30 percent of their length, measured at the lowest vocal pitch examined.

Measurements of vocal-fold length by means of laryngeal photography are presented and discussed in Hollien (1960) and Hollien and Moore (1960). Vocal-fold length was observed to increase systematically with increases in vocal pitch, while no uniform pattern of elongation or shortening was observed in the falsetto register. In some cases, the vocal folds were shorter in falsetto than in high-chest voice. The authors also noted that the vocal folds in abducted position were longer than in phonation.

Variations in the mass of the vocal folds participating in vibratory motion can be estimated by measuring the vocal-fold thickness during phonation. For this, laryngeal laminagraphic studies have been reported by Hollien and Curtis (1960), Hollien (1962), and Hollien and Colton (1969). Their results revealed that vocal-fold thickness decreased systematically with increase in vocal pitch in the modal (chest) register but did not do so in the major portion of the falsetto register. Vocal-fold thickness in falsetto appeared to fluctuate slightly around a level at or near the smallest value in the modal register.

Elongation of the vocal folds is achieved by the contraction of the cricothyroid muscle, which has the effect of tilting the thyroid cartilage forward relative to the cricoid. Isshiki (1959), with electrical stimulation of the larynx of the dog, found that vocal pitch rose to some extent with increasing stimulation of the recurrent nerves, while it showed marked elevation with additional stimulation of the external branch of the superior laryngeal nerves which innervate cricothyroid muscles. Rubin (1963), in similar experiments, also reported that the contraction of the thyroarytenoids alone raised pitch but to a much lesser degree than did the contraction of the cricothyroid alone.

Activities of the intrinsic muscles of human larynges have been studied electromyographically by many investigators. Katsuki (1950), Faaborg-Andersen (1957, 1965), Sawashima et al. (1958), Arnold (1961), and Faaborg-Andersen et al. (1967) have all reported an increase in the electrical activity of the cricothyroid muscle associated with an increase in fundamental frequency of chest voice. A similar activity in the vocalis muscle has been reported by Faaborg-Andersen (1957, 1965) and Sawashima et al. (1958). Hirano et al. (1969, 1970) claimed that the activities of the cricothyroid, the vocalis, and the lateral cricoarytenoid muscles were positively related to fundamental frequency of voice. Using averaged EMG data, Sawashima et al. (1969) noted that variation in the activity of the cricothyroid muscle has a more linear relationship to the fundamental frequency than has the vocalis muscle. Activities of the intrinsic muscles in falsetto are somewhat different from those in chest register. For rising vocal pitch with a shift in register from chest to falsetto, Faaborg-Andersen noted less increase in the activity of the cricothyroid muscle than was found within the chest register. Similar results were obtained by Hirano et al. Less activity of the cricothyroid muscle in falsetto compared to chest voice was observed by Sawashima et al. Faaborg-Andersen noted little or no increase in vocalis muscle activity for the increase of vocal pitch with the shift in register. Hirano et al. (1969, 1970) and Sawashima et al. (1958) have reported a much smaller degree of activity of the vocalis muscle in falsetto than in chest voice.

The experimental results mentioned above reveal that, in chest register, pitch rise is achieved by contraction of the vocalis muscle in combination with its elongation caused by contraction of the cricothyroid muscle, whereas, in falsetto, a different type of laryngeal adjustment is exerted and the participation of the vocalis muscle is very slight.

An increase in the activity of the lateral cricoarytenoid muscle for higher fundamental frequency in chest register, which has been reported by Hirano et al. (1969, 1970) compares with the effect of the medial compression in the above mentioned experiment of van den Berg and Tan (1959).

The contribution of the extrinsic laryngeal muscles to vocal pitch control has been discussed by several investigators. Sonninen (1956) claimed there was a forward pull of the thyroid cartilage (vocal-fold elongation) by simultaneous contraction of the sternohyoid, hyothyroid, and the ventral group of the suprahyoid muscles, the cricoid cartilage being fixed in position by the cricopharyngeal muscle. Zenker and Zenker (1960) emphasized the effect of the thyropharyngeal muscle which approximates the plates of the thyroid cartilage and displaces the anterior origin of the vocal folds in a forward direction.

Active pitch-lowering mechanisms in the larynx are still to be examined. Lindqvist (1969) proposed a pitch-lowering mechanism by a *laryngealization*, i.e., contraction of the aryepiglottic sphincter, for shortening the vocal folds. Among the extrinsic muscles, the sternothyroid muscle was assumed, by some investigators, to have a pitch-lowering effect by tilting the thyroid cartilage backward. However, Sonninen (1956) has shown that the contraction of the sternothyroid could cause tilting of the cartilage in either direction depending on the position of the head and the spine. Zenker and Zenker (1960) claimed the cricopharyngeal muscle to be of importance for shortening the vocal folds by pulling the cricoid cartilage backward and upward. They also proposed a functional chain of arytenoid cartilage, aryepiglottic muscle, epiglottis, tongue, hyoid bone, and mandible as a device for shortening the vocal folds. Minnigerode (1967) disagreed with Zenker and Zenker and claimed that the function of the cricopharyngeal muscle would be rather synergic or auxiliary to the cricothyroid muscle since it would assist a horizontal relative displacement of the thyroid and cricoid cartilages.

Electromyographic studies of the extrinsic muscles have presented rather complex data on their activities in connection with variations in vocal pitch. Faaborg-Andersen and Sonninen (1960) examined the sternothyroid, the mylohyoid, and the thyrohyoid muscles. Pronounced activity in the sternothyroid muscle was observed during phonation at low and high pitches, while there was a decrease in the activity in the middle of the pitch range. The mylohyoid muscle showed pronounced activity only in the middle of the tone scale, and the activity of the thyrohyoid muscle increased during middle- and high-pitched phonation. Arnold (1961) reported increasing activity in the sternothyroid muscle along the ascending scale of vocal pitch. Zenker and Zenker (1960) noted that activities of the extrinsic muscles were found to be minimal during phonation at the mean of conversational pitch, which is close to the rest position of the larynx. Any effort to change vocal pitch from this level showed an increase in activity of the muscles. Thus, pronounced activity of the thyropharyngeal muscle associated with some increase in cricopharyngeal activity was observed for raising pitch, and a predominant increase in activity of the cricopharyngeal muscle accompanied by some increase in thyropharyngeal activity for lowering pitch. Activity of the sternohyoid muscle was studied by Hirano et al. (1967). The activity was prominent in both high and low pitches, while it was less marked in the middle portion of the pitch range.

The data mentioned above indicate that there is no simple relationship to be found between vocal pitch and the action of the extrinsic laryngeal muscles. Complex patterns of their activities are also considerably influenced by the position of the head, the larynx, the lower jaw, and the tongue.

Intensity of voice is considered to be regulated by subglottic pressure with or without active participation of laryngeal muscles. Rubin (1963), in his experiments on the dog, reported a rise in sound-pressure levels in response both to increased air flow and to increased contraction of the thyroarytenoid and cricothyroid muscles.

Isshiki (1964), measuring the subglottic pressure and the air-flow rate in relation to vocal intensity in human subjects, concluded that, in the very low pitches, the air-flow resistance at the glottis, which reflects laryngeal control, dominantly correlates with the intensity variation, the correlation becoming less apparent as the pitch is raised, until the extremely high pitch range voice-intensity variation is correlated with air-flow rate.

Adjustments of the glottal resistance must be reflected in the electrical activities of the laryngeal muscles. Electromyographic studies have shown somewhat controversial data on the laryngeal control of vocal intensity. No significant change in activity of either the cricothyroid or the vocalis muscle for varying vocal intensity was observed by Faaborg-Andersen (1957, 1965) and Sawashima et al. (1958), while Arnold (1961) did find greater cricothyroid activity for higher intensities. Greater activity in the vocalis, arytenoid, and posterior cricoarytenoid muscles for increased vocal intensity was reported by Zboril (1965). Hirano et al. (1969, 1970) observed increased activity in the vocalis and lateral cricoarytenoid muscles for higher intensities in low-chest voice, while in falsetto the activity remained unchanged or decreased with increasing intensity. Cricothyroid activity usually remained unchanged or decreased. This result is consistent with the aerodynamic data obtained by Isshiki. The decrease in the muscle activity seems to reflect compensatory laryngeal adjustments for maintaining constant pitch, which otherwise tends to be raised by increased air flow.

From the data mentioned above, it is difficult to obtain clear conclusions on vocal-intensity control. Laryngeal adjustments for intensity control seem less pronounced than for pitch control, and intensity control is probably closely interrelated with respiratory and/or aerodynamic conditions.

Laryngeal Adjustments During Speech Articulation

Experimental data directly relevant to laryngeal behavior in speech utterances is, at present, far less readily available than that for sustained phonation. This may be attributed in part to the technical difficulty in obtaining reliable data without disturbing natural movements of articulatory organs. Furthermore, experimental conditions in speech can not be simplified or artificially controlled as in sustained phonation.

Electromyographic studies of the laryngeal muscles in relation to intonation or accent patterns in speech have been reported by several investigators.

For a pitch rise at the end of utterances in an interrogative form, an increase of electrical activity was observed in the cricothyroid and the vocalis muscles (Hirano et al., 1969; Harris et al., 1969) and also in the lateral cricoarytenoid muscle (Hirano et al., 1969). Hirano et al. (1969) reported an increase in activity of the cricothyroid, the vocalis, and the lateral cricoarytenoid muscles associated with an accentuation put on words in English sentences. Simada and Hirose (1970), in utterances of Japanese words with different pitch accent patterns, observed increased activity in the cricothyroid and the lateral cricoarytenoid muscles in correspondence to the accent feature. The cricothyroid muscle activity, in particular, showed almost clear-cut correspondence to the time course of the fundamental frequency. Similar results have been shown in EMG of the cricothyroid and the vocalis muscle in relation to Swedish accent patterns (Gårding et al., 1970).

Beside vocal pitch and intensity control, an important participation of laryngeal control in speech is seen in actualization of the voiced/voiceless distinction of consonants. Glottal conditions during consonant articulations have been studied by several investigators.

Transillumination data were reported by Slis and Damsté (1967) and Lisker et al. (1969). The results of Slis and Damsté were:

1) The glottis was open for voiceless consonants, the degree of opening being equal to the maximum opening in the vibratory cycles of the adjoining vowels for plosives and about twice as large for fricatives.

2) The glottal condition for the voiced plosives and fricatives was almost the same as for the adjoining vowels with some exceptions where the glottis was found to be open.

3) The laryngeal behavior for the glides, the nasals, and the liquids did not differ from that for the adjoining vowels.

4) In intervocalis /h/, the vocal folds were vibrating while the degree of glottal opening equalled that of the voiceless fricatives.

5) In whispered speech, the glottal opening was greater for voiceless consonants than for their voiced counterparts.

Results of Lisker et al. for running speech with American English were as follows:

1) Voiceless stops were produced with either opening of the glottis or interruption of voicing. Some voiceless stops in unstressed positions were produced with a closed glottis or uninterrupted voicing.

2) Voiceless fricatives were produced with both opening of the glottis and interruption of voicing.

3) Voiced stops were mostly produced with a closed glottis and uninterrupted voicing.

4) Voiced fricatives were produced with either an open or a closed glottis, the voicing being uninterrupted in both cases.

Visual examination of the glottis by use of a fiberoptics has been reported by the present writer. Findings on Japanese syllables (Sawashima, 1968; Sawashima et al., 1968) and in running speech with American English (Sawashima et al., in press) are basically consistent with those obtained by transillumination mentioned above. Some additional data, including those obtained very recently, may be briefly mentioned here. With voiceless stops, the degree of glottal opening for the same phoneme varies considerably depending on the phonological environment and also on the speed of utterance. In some voiceless stops, the arytenoid cartilages appear to stay in the phonatory position with a narrow spindle-shaped opening along the membranous portion of the glottis. In English, the glottal openings for the aspirated voiceless stops are larger than for unaspirated ones. The relative timing between laryngeal and supraglottal articulatory gestures for voiceless stops was examined. For voiceless stops, the separation of the arytenoid cartilages shows a relative timing that varies considerably, occurring both before and after oral closure, with some intersubject differences. The cessation of glottal vibration occurs after the beginning of oral closure and the separation of the arytenoids. For unaspirated voiceless stops, the resumption of vibration takes place just at or immediately after the stop release while the arytenoid closure is completed shortly after the resumption of vibration. For aspirated voiceless stops, both the resumption of vibration and closure of the arytenoids take place much later than for nonaspirates.

These articulatory maneuvers, including the relative timing of laryngeal control and vocal tract control, are considered to be appreciably different from language to language. Some other languages with different manner characterizations of consonants have been studied. Vencov (1968), using a special device to make artificial releases during stop closure, claimed that, in Russian voiceless stops, the glottis stays in a position close enough to vibrate whenever transglottal air flow is resumed. Glottal openings for three types of Korean stops (unaspirated, slightly aspirated, heavily aspirated) were examined cineradiographically by Kim (1970). The results revealed a larger opening of the glottis at the time of release for the sounds with greater degrees of aspiration.

An electromyographic study on the glottal adjustments for consonant articulation was reported by Hiroto et al. (1967). During the articulation of Japanese VCV syllables, a temporary decrease in electrical activity in the adductor muscles (thyroarytenoid, lateral cricoarytenoid, and interarytenoid) and an increase in the activity of the abductor muscle (posterior cricoarytenoid) were found corresponding to voiceless consonants, while no change was detectable in the cricothyroid muscle. No differences in the EMG patterns were observed between voiced consonants and vowels. The results indicate an opening gesture of the glottis for voiceless consonants, where the cricothyroid muscle does not participate.

As mentioned in the beginning of this section, experimental research on laryngeal adjustments during speech utterances is now still in the preliminary stages. Data obtained so far seem to indicate complexity and variability of the laryngeal maneuvers in actual speech as shown in the articulatory gestures in the supraglottal organs. Further data collection by means of various modern research techniques dealing with different physical quantities, including for example aerodynamic measures, is necessary for the description of the laryngeal articulatory mechanism.

REFERENCES

Abo-El-Enein, M. A. and B. Wyke. 1966a. Myotatic reflex system in the intrinsic muscles of the larynx. J. Anat. 100.926-927.

Abo-El-Enein, M. A. and B. Wyke. 1966b. Laryngeal myotatic reflexes. Nature 209.682-686.

Arnold, G. E. 1961. Physiology and pathology of the cricothyroid muscle. Laryngoscope 71.687-753.

Asano, H. 1968. Application of the ultrasonic pulse-method on the larynx. Jap. J. Otol. Tokyo 71.895-916 (in Japanese).

Baken, J. R. 1969. Distribution of neuromuscular spindles in intrinsic muscles of a human larynx. Ph.D. thesis, Columbia Univ.

Basmajian, J. V. and C. R. Dutta. 1961. Electromyography of the pharyngeal constrictors and levator palatini in man. Anat. Record. 139.561-563.

Basmajian, J. V. and G. Stecko. 1962. A new bipolar electrode for electromyography. J. Appl. Physiol. 17.849.

Beach, J. L. and C. A. Kelsey. 1969. Ultrasonic Doppler monitoring of vocal fold velocity and displacement. JASA. 46.1045-1047.

Berendes, J. 1964. Die Leistung der Kehlkopfmuskulatur unter dem Aspekt der Electronenmikroskopie und der Enzymchemie. Mschr. Ohrenheilk. 98.524-528.

Berendes, J. and W. Vogell. 1960. Kehlkopfmuskeln im elektronenmikroskopischen Bild. Arch. Ohr-Nas-u. Kehlk. Heilk. 176.730-735.

Berg, Jw. van den. 1960. Vocal ligaments versus registers. Current Problems in Phoniatrics and Logopedics 1.19-34.

Berg, Jw. van den. 1962. Modern research in experimental phoniatrics. Folia phoniat. 14.81-149.

Berg, Jw. van den and T. S. Tan. 1959. Results of experiments with human larynxes. Prakt. oto-rhino-laryng. 21.425-450.

Bianconi, R. and G. Molinari. 1962. Electromyographic evidence of muscle spindles and other sensory endings in the intrinsic laryngeal muscles of the cat. Acta. oto-laryng. 55.253-259.

Bowden, R. E. M. and J. L. Sheuer. 1960. Weights of abductor and adductor muscles of the human larynx. J. Laryng. 74.971-980.

Buchthal, F., C. Guld, and R. Rosenfalck. 1957. Multielectrode study of the tensity of a motor unit. Acta Physiol. Scand. 39.83-104.

Colman, F. R. and R. W. Wendahl. 1968. On the validity of laryngeal photosensor monitoring. JASA. 44.1733-1735.

Cooper, S. and J.C. Eccles. 1930. The isometric responses of mammarian muscles. J. Physiol. 69.376-85.

Damsté, P. H., H. Hollien, P. Moore, and Th. Murry. 1968. An X-ray study of vocal fold length. Folia phoniat. 20.349-359.

Dedo, H. and E. Dunker. 1966. The volume conduction of motor unit potentials. Electroenceph. clin. Neurophysiol. 20.608-613.

Dedo, H. and W. H. Hall. 1969. Electrodes in laryngeal electromyography: Reliability comparison. Ann. Otol. 78.172-180.

Dunker, E. 1968. The central control of laryngeal function. Ann. New York Acad. Sciences 155.112-121.

Dunker, E. 1969. Neue Ergebnisse der Kehlkopfphysiologie. Folia phoniat. 21.161-178.

Esslen, E. and B. Schlosshauer. 1960. Zur Reflexinnervation der inneren Kehlkopfmuskeln. Folia phoniat. 12.161-169.

Faaborg-Andersen, K. 1957. Electromyographic investigation of intrinsic laryngeal muscles in humans. Acta Physiol. Scand. 41.Suppl. 140.

114

Faaborg-Andersen, K. 1965. Electromyography of laryngeal muscles in humans. Technics and results. S. Karger, A.G. Basel.

Faaborg-Andersen, K. and A. Sonninen. 1960. The function of the extrinsic laryngeal muscles at different pitches. An electromyographic and roentgenologic investigation. Acta oto-laryng. 51.89-93.

Faaborg-Andersen, K., N. Yanagihara, and H. von Leden. 1967. Vocal pitch and intensity regulation. A comparative study of electrical activity in the cricothyroid muscle and the airflow rate. Arch. Otolaryng. 85.448-454.

Fabre, M. P. 1957. Un procédé electrique percutané d'inscription de l'accolement glottique au cours de la phonation: Glottographie de haute fréquence. Premiers résultats. Bull. Acad. nat. Méd. 141.66-69.

Fabre, M. P. 1958. Etude comparée des glottogrammes et des phonogrammes de la voix humaine. Ann. d'otolaryng. 75.767-775.

Fabre, M. P. 1961. Glottographie respiratoire: Appareillage et premiers résultats. C. R. Acad. Sci. Paris 252.1386-1388.

Fant, G., J. Ondráčková, J. Lindqvist, and B. Sonesson. 1966. Electrical glottography. STL-QPSR-4/1966 15-21.

Fischer-Jørgensen, E., B. Frøkjaer-Jensen, and J. Rischel. 1966. Preliminary experiments with the Fabre glottograph. Annual Report of the Institute of Phonetics of Univ. Copenhagen 1.22-29.

Frable, M. A. 1961. Computation of motion of cricoarytenoid joint. Arch. Otolaryng. 73.551-556.

Frøkjaer-Jensen, B. 1967. A photo-electric glottograph. Annual Report of the Institute of Phonetics of Univ. Copenhagen 2.5-19.

Frøkjaer-Jensen, B. 1968. Comparison between a Fabre glottograph and a photo-electric glottograph. Annual Report of the Institute of Phonetics of Univ. Copenhagen 3.9-16.

Frøkjaer-Jensen and P. Thorvaldsen. 1968. Construction of a Fabre glottograph. Annual Report of the Institute Phonetics of Univ. Copenhagen 3.1-8.

Ganz, H. 1962. Untersuchungen zur capillären Blutversorgung des Musculus vocalis beim Menschen. Arch. Ohr-Nas-u. Kehlk.Heilk. 179.338-350.

Ganz, H. 1964. Enzymmuster von Kehlkopfmuskeln des Menschen und ihre Bedeutung für deren Funktion an der Glottis. Arch. Ohr-Nas-u. Kehlk. Heilk. 184.1-10.

Gårding, E., O. Fujimura, and H. Hirose. 1970. Laryngeal control of Swedish word tones: A preliminary report on an EMG study. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, Univ. Tokyo) No. 4. 45-54.

Gerebtzoff, M. A. and G. Lepage. 1960. Cited from Hiroto (1966).

Goerttler, K. 1950. Die Anordnung, Histologie und Histogenese der quergestreiften Muskulatur im menschlichen Stimmband. Zschr. Anat. Entw. 115.352-401.

Gordon, G. and C.G. Phillips. 1953. Slow and rapid components in a flexor muscle. Quart. J. Exp. Physiol. 38.35-45.

Gracheva, M. S. 1963. Sensory innervation of locomotor apparatus of the larynx. Fed. Proc. 22.T1120-T1123.

Greiner, G. F., F. Isch, and J-C Lafon. 1958. A propos de quelques cas d'électromyographie de la corde vocal chez l'homme. Ann. d'otolaryng. 75.23-30.

Harris, K. S., T. Gay, P. Lieberman, and G. Sholes. 1969. The function of muscles in control of stress and intonation. Status Report on Speech Research, SR-19/20 127-138. Haskins Laboratories.

Hast, M. H. 1966a. Mechanical properties of the cricothyroid muscle. Laryngoscope 76.537-548.

Hast, M. H. 1966b. Physiological mechanism of phonation. Tension of the vocal fold muscle. Acta oto-laryng. 62.309-318.

Hast, M. H. 1967. Mechanical properties of the vocal fold muscle. Pract. Oto-Rhino-Laryng. 29.53-56.

Hast, M. H. 1968. Studies on the extrinsic laryngeal muscles. Arch. Otolaryng. 88.273-278.

Hast, M. H. 1969. The primate larynx. A comparative physiological study of intrinsic muscles. Acta oto-laryng. 67.84-92.

Hirano, M., Y. Koike, and H. von Leden. 1967. The sternohyoid muscle during phonation. Electromyographic studies. Acta oto-laryng. 64.500-507.

Hirano, M. and J. Ohala. 1969. Use of hooked-wire electrodes for electromyography of the intrinsic laryngeal muscles. JSHR. 12.362-373.

Hirano, M., J. Ohala, and W. Vennard. 1969. The function of the laryngeal muscles in regulating fundamental frequency and intensity of phonation. JSHR. 12.616-628.

Hirano, M., W. Vennard, and J. Ohala. 1970. Regulation of register, pitch and intensity of voice: An electromyographic investigation of intrinsic laryngeal muscles. Folia phoniat. 22.1-20.

Hirose, H. 1961. Afferent impulses in the recurrent laryngeal nerve in the cat. Laryngoscope 71.1196-1206.

Hirose, H., T. Ushijima, T. Kobayashi, and M. Sawashima. 1969. An experimental study of the contraction properties of the laryngeal muscles in the cat. Ann. Otol. 78.297-307.

Hiroto, I. 1966. Pathophysiology of the larynx. From the aspect of the mechanism of phonation. Pract. Otol. Kyoto 39, Suppl. 1. 229-291 (in Japanese).

Hiroto, I., M. Hirano, Y. Toyozumi, and T. Shin. 1962. A new method of placement of a needle electrode in the intrinsic laryngeal muscles for electromyography: Insertion through the skin. Oto-rhino-laryng. Clinic, Kyoto 55.499-504 (in Japanese).

Hiroto, I., M. Hirano, T. Toyozumi, and T. Shin. 1967. Electromyographic investigation of the intrinsic laryngeal muscles related to speech sounds. Ann. Otol. 76.861-872.

Hollien, H. 1960. Vocal pitch variation related to changes in vocal fold length. JSHR. 3.150-156.

Hollien, H. 1962. Vocal fold thickness and fundamental frequency of phonation. JSHR. 5.237-243.

Hollien, H. and R. H. Colton. 1969. Four laminagraphic studies of vocal fold thickness. Folia phoniat. 21.179-198.

Hollien, H. and J. F. Curtis. 1960. A laminagraphic study of vocal pitch. JSHR. 3.361-370.

Hollien, H. and G. P. Moore. 1960. Measurements of the vocal folds during changes in pitch. JSHR. 3.157-165.

Isshiki, N. 1959. Regulatory mechanism of the pitch and volume of voice. Oto-rhino-laryng. Clinic, Kyoto 52.1065-1094.

Isshiki, N. 1964. Regulatory mechanism of voice intensity variation. JSHR. 7.17-29.

Katsuki, Y. 1950. The function of the phonatory muscles. Jap. J. Physiol. 1.29-36.

Kawano, A. 1968. Biochemical and electronmicroscopic investigation of the laryngeal muscles. Otologia Fukuoka 14. Suppl. 2 (in Japanese).

Kelsey, C. A., R. J. Woodhouse, and F. D. Minifie. 1969. Ultrasonic observations of coarticulation in the pharynx. JASA. 46.1016-1018.

Kim, C. W. 1970. A theory of aspiration. Phonetica 21.107-116.

Kirchner, J. A. and B. Wyke. 1964a. Laryngeal articular reflexes. Nature 202.600.

Kirchner, J. A. and B. Wyke. 1964b. Electromyographic analysis of laryngeal articular reflexes. Nature 203.1243-1245.

Kirchner, J. A. and B. Wyke. 1965. Articular reflex mechanism in the larynx. Ann. Otol. 74.749-768.

Klatt, D. H., K. N. Stevens, and J. Mead. 1968. Studies of articulatory activity and airflow during speech. Ann. New York Acad. Sciences. 155.42-54.

König, W. F. and H. von Leden. 1961. The peripheral nervous system of the human larynx. Arch. Otolaryng. 74.153-163.

Lindqvist, J. 1969. Laryngeal mechanisms in speech. STL-QPSR 2-3/1969 26-31.

Lisker, L., A. S. Abramson, F. S. Cooper, and M. H. Schvey. 1969. Transillumination of the larynx in running speech. JASA. 45.1544-1546.

Lucas Keen, M. F. 1961. Muscle spindles in human laryngeal muscles. J. Anat. 95.25-29.

Malécot, A. and K. Peebles. 1965. An optical device for recording glottal adduction-abduction during normal speech. Zphon 18.545-550.

Mårtensson, A. 1963. Reflex responses and recurrent discharges evoked by stimulation of laryngeal nerves. Acta Physiol. Scand. 57.248-269.

Mårtensson, A. 1964. Proprioceptive impulse patterns during contraction of intrinsic laryngeal muscles. Acta Physiol. Scand. 62.176-194.

Mårtensson, A. 1967. Recurrent discharges studied in single units of laryngeal muscles. Acta Physiol. Scand. 70.221-228.

Mårtensson, A. 1968. The functional organization of the intrinsic laryngeal muscles. New York Acad. Sciences. 155.91-97.

Mårtensson, A. and C. R. Skoglund. 1964. Contraction properties of intrinsic laryngeal muscles. Acta Physiol. Scand. 60.318-336.

Matzelt, D. and K. H. Vosteen. 1963. Elektronenoptische und enzymatische Untersuchungen an menschlicher Kehlkopfmuskulatur. Arch. Ohr-Nas-u. Kehlk.Heilk. 181.447-457.

Michel, Cl. van. and L. Raskin. 1969. L'électroglottomètre Mark 4, son principe, ses possibilités. Folia phoniat. 21.145-157.

Minifie, F. D., C. A. Kelsey, and T. J. Hixon. 1968. Measurement of vocal fold motion using an ultrasonic Doppler velocity monitor. JASA. 43.1165-1169.

Minnigerode, B. 1967. Untersuchungen zur Bedeutung der extralaryngealen Muskulatur für den Phonationsakt unter besondere Berücksichtigung des Musculus cricopharyngeus. Arch. klin. exper. Ohren-Nasen-u. Kehlkopfheilk. 188.604-623.

Miura, T. 1969. Mode of vocal cord vibration: A study with ultrasonoglottography. Jap. J. Otol. Tokyo. 72.985-1002 (in Japanese).

Nakamura, A. 1965. Histoanatomical studies on the human vocal cords. Jap. J. Otol. Tokyo. 68.126-136 (in Japanese).

Negus, V. E. 1962. The comparative anatomy and physiology of the larynx. Hafner Publ. Co., New York.

Ohala, J. 1966. A new photo-electric glottograph. Working Papers in Phonetics. UCLA. No. 4. 40-52.

Paulsen, K. 1958. Über den Bedeutung von Muskelspindeln und Schreimhautreceptoren bei der Phonation. Arch. Ohr-Nas-u. Kehlk.Heilk. 173.500-503.

Rubin, H. J. 1963. Experimental studies on vocal pitch and intensity in phonation. Laryngoscope 73.973-1015.

Rubin, H. J. and C. C. Hirt. 1960. The falsetto: A high speed cinematographic study. Laryngoscope 70.1305-1324.

Rubin, H. J., M. Le Cover, and W. Vennard. 1967. Vocal intensity, subglottal pressure and airflow relationships in singers. Folia phoniat. 19.393-413.

Rudolph, G. 1961. Spiral nerve-endings (proprioceptors) in the human vocal muscle. Nature 190.726-727.

Sampson, S. and C. Eyzaguirre. 1964. Some functional characteristics of mechanoreceptors in the larynx of the cat. J. Neurophysiol. 27. 464-480.

Sawashima, M. 1968. Movements of the larynx in articulation of Japanese consonants. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, Univ. Tokyo) No. 2. 11-20.

Sawashima, M., A. S. Abramson, F. S. Cooper, and L. Lisker. In press. Observing laryngeal adjustments during running speech by use of a fiberoptics system. To appear in Phonetica.

Sawashima, M., T. Gay, and K. S. Harris. 1969. Laryngeal muscle activity during vocal pitch and intensity changes. Status Report on Speech Research SR-19/20. 211-220 Haskins Laboratories.

Sawashima, M. and H. Hirose. 1968. New laryngoscopic technique by use of fiber optics. JASA. 43.168-169.

Sawashima, M., H. Hirose, S. Kiritani, and O. Fujimura. 1968. Articulatory movements of the larynx. Proc. 6th Internat. Congr. Acoust. B-1-B-4.

Sawashima, M., M. Sato, S. Funasaka, and G. Totsuka. 1958. Electromyographic study of the human larynx and its clinical application. Jap. J. Otol. Tokyo. 61.1357-1364 (in Japanese).

Schlosshauer, B. and K. H. Vosteen. 1957. Über die Anordnung und Wirkungsweise der im Conus elasticus ansetzenden Fasern des Stimmuskels. Zschr. Laryng. Rhin. Otol. 36.642-650.

Shipp, T., W. W. Deatsch, and K. Robertson. 1968. A technique for electromyographic assessment of deep neck muscle activity. Laryngoscope 78.418-432.

Simada, Z. and H. Hirose. 1970. The function of the laryngeal muscles in respect to the word accent distinction. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, Univ. Tokyo) No. 4. 27-40.

Slis, J. H. and P. H. Damsté. 1967. Transillumination of the glottis during voiced and voiceless consonants. IPO Annual Progress Report No. 2. 103-109 Eindhoven, Holland.

Sonesson, B. 1958. Die funktionelle Anatomie des Cricoarytenoidgelenkes. Zschr. Ant. Entw. 121.292-303.

Sonesson, B. 1960. On the anatomy and vibratory pattern of the human vocal folds. Acta oto-laryng. Suppl. 156.

Sonninen, A. A. 1956. The role of the external laryngeal muscles in length-adjustment of the vocal cords in singing. Acta oto-laryng. Suppl. 130.

Sonninen, A. A. 1968. The external frame function in the control of pitch in the human voice. Ann. New York Acad. Sciences. 155.68-89.

Suzuki, M. and J. A. Kirchner. 1968. Afferent nerve fibers in the external branch of the superior laryngeal nerve in the cat. Ann. Oto. 77. 1059-1070.

Suzuki, M. and J. A. Kirchner. 1969. Sensory fibers in the recurrent laryngeal nerve: An electrophysiological study of some laryngeal afferent fibers in the recurrent laryngeal nerve of the cat. Ann. Otol. 78.21-31.

Takase, S. 1964. Studies on the intrinsic laryngeal muscles of mammals: Comparative anatomy and physiology. Otologia Fukuoka 10. Suppl. 1 (in Japanese).

Tomita, H., Y. Koike, and S. Tomita. 1967. A histochemical study on the intrinsic laryngeal muscles. Otologia Fukuoka 13.286-294 (in Japanese).

Vencov, A. V. 1968. A mechanism for production of voiced and voiceless intervocalic consonants. Zphon. 21.140-144.

Von Leden, H. and P. Moore. 1961. The mechanism of the cricoarytenoid joint. Arch. Otolaryng. 73.541-550.

Weiss, O. 1914. Cited from Sonesson (1960).

Wustrow, F. 1952. Bau und Funktion des menschlichen Musculus vocalis. Zschr. Anat. Entw. 116.506-522.

Wyke, B. 1967. Recent advances in the neurology of phonation: Phonatory reflex mechanisms in the larynx. British J. Disorders of Communication. 2.2-14.

Zboril, M. 1965. Electromyographie der inneren Kehlkopfmuskeln bei verschiedenen Phonationstypen. Arch. Ohr-Nas-u. Kehlk.Heilk. 184.443-449.

Zenker, W. 1964. Vocal muscle fibers and their motor endplates. In Research Potentials in Voice Physiology. D. W. Brewer, Ed. State Univ. of New York.

Zenker, W. and A. Zenker. 1960. Über die Regerung der Stimlippenspannung durch von aussen eingreifende Mechanismen. Folia phoniat. 12.1-36.

Speech Synthesis for Phonetic and Phonological Models*

Ignatius G. Mattingly[+]
Haskins Laboratories, New Haven

INTRODUCTION

The linguist today, and more especially the phonologist, is aware of two very active areas of investigation. One of them, generative phonology, is at the very center of his field of vision and cannot be ignored. The other, experimental phonetics, may seem less directly relevant to his concerns, even though he accepts the truism that phonology rests ultimately on a phonetic basis. As modern experimental phonetics grows more technical and is increasingly dominated by the psychologist, the physiologist, the electronics engineer, and the speech scientist, the phonologist is more likely than ever to be put off and to yield to the temptation to do phonology on the basis of phonetic folklore.

In this situation, it is fortunate that there is a group of investigators whose efforts tend to bridge this gap: those engaged in synthesis of speech by rule. By synthesis by rule, we mean the automatic production of audible synthetic speech from a symbolic transcription by a process that models phonetic and phonological rules in some nontrivial way. With the help of modern electronic technology, it is now perfectly possible to type a transcription on a computer typewriter and immediately hear the corresponding utterance. But the investigator who undertakes synthesis by rule not only makes use of advanced technical facilities, he also attempts, of necessity, to integrate and generalize into a system those findings of experimental phonetics that are relevant to phonology. On this account, at least, synthesis by rule should intrigue the phonologist.

We believe that there are other reasons as well why the phonologist should be interested, and we shall try to make them clear below. We shall also say a little about the techniques of speech synthesis, give some account of the development of synthesis by rule, describe a number of current approaches to the task, and finally, suggest some possible directions that this work may take in the future.

Speech synthesized by rule is not the only kind of synthetic speech.[1]

---

*Chapter prepared for Current Trends in Linguistics, Vol. XII, Thomas A. Sebeok, Ed. (The Hague: Mouton).

[+] Also, University of Connecticut, Storrs

[1] For general discussions of speech synthesis, see Wheatstone (1837); Dudley and Tarnoczy (1950); Fant (1958); Cooper (1962); and Flanagan (1965:167-191).

There are several others which should be mentioned, if only because investigations motivated by these uses have led to technical advances of general benefit. Thus, much of the speech synthesis research of the past thirty years has been prompted by interest in vocoding (i.e., voice coding). The channel capacity (equivalently, the bandwidth in the radio spectrum) required for transmission of speech is many times greater than it ought to be, considering the amount of information, in Shannon's sense, that is carried by the speech signal. Since the channel capacity available for radio and cable communications is limited, many schemes have been devised to "compress" speech by analyzing the speech wave and transmitting only the information needed to synthesize an intelligible version at the receiving end. For example, in Dudley's (1939) original Vocoder, built at Bell Telephone Laboratories, the spectrum of telephone speech (250-3000 Hz) is analyzed by a bank of 10 filters. The smoothed, rectified output of each filter represents the energy in a certain part of the spectrum as a function of time. Another circuit tracks F_0, the fundamental frequency (for voiceless excitation, the output of this circuit is zero). The Vocoder transmits the outputs of the F_0 tracker and of the filters. Since these functions vary relatively slowly, the channel capacity needed for all 11 functions is far less than the unprocessed speech signal would require. To synthesize the speech, the frequency of a buzz source is varied according to the F_0 function (a hiss source is used when this function has zero value). The buzz or hiss excites each of a set of filters matching those used in the analysis, and the amplitude of the output from each synthesizing filter is determined by the function for the corresponding analyzing filter. Summing the outputs of the synthesis filters yields an intelligible version of the original speech.

A second type of vocoder is the formant vocoder (Munson and Montgomery, 1950). In a formant vocoder, the analyzer tracks the excitation state, F_0, and the frequencies and amplitudes of the lowest three formants of the original speech and transmits these functions; in the synthesizer, resonant circuits representing the three formants are appropriately excited, and the transmitted functions also determine the frequency and the amplitude for each resonator. The saving in channel capacity is greater than for a filter-bank vocoder, but correct analysis is much more difficult. Both filter-bank and formant synthesizers have proved to be of value for phonetic and phonological research as well as for communications.

Besides vocoding, there are certain other possible applications for synthetic speech. If it is necessary for a machine to communicate with its user--a computer operator or a student undergoing computer-assisted instruction--and heavy demands are already being made on his visual attention, spoken messages may be the solution. But if fast random access to a large inventory of messages is required, storage of natural speech becomes cumbersome, for speech makes the same exorbitant demands on storage capacity as it does on channel capacity (Atkinson and Wilson, 1968). Synthetic speech, if it could be stored in some kind of minimal representation, would be an attractive alternative. Still another application is a reading machine for the blind. In such a device, printed text must be converted to spoken output with the aid of a dictionary in which written and spoken elements are matched. If these elements are naturally spoken words, the output rate cannot exceed ordinary speaking rates without distortion and (on account of the difficulty of abutting the words closely without losing intelligibility) will actually be considerably slower. Yet the blind user would be happy with an output several times faster than natural speech. A potential solution is suggested by the fact that speech can be synthesized at two or three times normal speed without much loss of intelligibility (Cooper et al., 1969).

In addition to these practical applications, synthetic speech is used for psychological research into the nature of speech perception itself. Synthetic stimuli can be produced which are simple and closely controlled and which, in some cases, could not have been produced by a human speaker at all. Such stimuli have been used to study categorical and continuous perception of speech sounds (Liberman et al., 1957), differences between perception of speech and nonspeech signals (Liberman et al., 1961), and hemispheric localization of speech processing (Shankweiler and Studdert-Kennedy, 1967). These investigations, apart from their intrinsic interest, are an essential preliminary to synthesis by rule.

Another research application of synthesis is the close imitation of natural speech utterances. It is of some interest to know just how faithfully a particular synthesizer can simulate natural speech, without any assumptions being made about the structure of speech or language beyond those built into the synthesizer. Such investigations explore the limitations of the synthesizer. If the best imitations that could be achieved in this way were indeed quite poor, this fact would discourage any endeavor making use of synthetic speech. Fortunately, at least one investigator, Holmes (1961; see also Holmes et al., 1964), using a formant synthesizer, has been able to synthesize sentences that are extremely natural and virtually impossible to distinguish from their originals. This is good evidence that progress in other applications of speech synthesis is, at any rate, not limited by the quality of the synthesizer.

HISTORICAL DEVELOPMENT OF SYNTHESIS-BY-RULE TECHNIQUES

The applications we have just summarized are quite recent. The traditional motivation for research in speech synthesis has been simply to explain how man used his vocal tract to produce connected speech. In a broad sense, such research is synthesis by rule, though it was a long time before the notion of a rule became obvious and the importance of an explicit formulation of the rules was recognized.

The idea of an artificial speaker is very old, an aspect of man's longstanding fascination with humanoid automata. Gerbert (d. 1003), Albertus Magnus (1198-1280), and Roger Bacon (1214-1294) are all said to have built speaking heads (Wheatstone, 1837). However, historically attested speech synthesis begins with Wolfgang von Kempelen (1734-1804), who published an account of his twenty years of research in 1791 (see also Dudley and Tarnoczy, 1950). Von Kempelen's synthesizer was a windbox driven by a bellows. One output of the windbox led to a reed, to simulate vocal-cord excitation; the reed was followed by a short neck and a bell-shaped, rubber mouth. Deforming the mouth changed the quality of the sound made by the reed. Projecting from the neck were two tubes which could be opened to make nasal sounds. Other outputs of the box included special passages for the fricatives [s] and [š]. There was also a lever to modulate the vibration of the reed for trilled [r] and an auxiliary bellows for aspiration of voiceless stops. A human operator played the synthesizer like a musical instrument; he pumped the bellows with his right arm, operated the various levers and tubes with his right hand, and manipulated the rubber mouth with his left hand.

Von Kempelen claims to have synthesized a number of short utterances in various languages ("Leopoldus Secundus," "vous êtes mon ami"). In 1923, Paget operated a copy of the synthesizer built by Wheatstone (1837) and was able to produce a few isolated words (Paget, 1930:19). But whatever the quality of the synthesis, one cannot fail to be impressed by the insights into the nature of speech production reflected in the design of the synthesizer and manifest in von Kempelen's monograph. He understood the basic relationship between the larynx and the supraglottal cavities and realized the special problems posed by nasals and fricatives. He understood also the importance of what Fant et al. (1963) were later to call "synthesis strategy": a set of techniques for producing the various classes of sounds which exploits the possibilities of a particular synthesizer and minimizes its limitations. (Thus, von Kempelen made an [f] by closing all regular outlets from the windbox and building up enough pressure inside to force air through the leaks in the box!) The obvious limitations of his work were the need to use a mechanical system, the acoustic properties of which were neither easily predictable nor readily alterable and the use of a human operator for dynamic control, with the consequence that the "rules" were not explicit but were rather part of the operator's art. Interestingly enough, the Abbé Mical, a contemporary of von Kempelen, is supposed to have built a synthesizer controlled by a pinned cylinder, such as is used in a music box. Wheatstone (1837:40-41) considered and dismissed the possibility of fitting his copy of von Kempelen's synthesizer with a control device of this sort:

> It would be a very easy matter to add either a keyboard or a pinned cylinder to De Kempelen's instrument, so as to make the syllables which it uttered follow, each with their proper accentuations and rests; but unless the articulations were themselves more perfect, it would not be worth the trouble and expense.

On the other hand, without a well-specified input, such as a pattern of pins on a cylinder, how can the performance of the synthesizer be systematically studied and improved?

During the century after von Kempelen and Mical, other manually operated, mechanical speech synthesizers were developed. Besides Wheatstone's copy of von Kempelen's synthesizer, there was, for instance, Faber's Euphonia (Gariel, 1879), which according to Dudley and Tarnoczy (1950) had variable pitch and sang "God Save the Queen."

But the next real step forward was Dudley's Voder (Dudley et al., 1939), an offshoot of his Vocoder, described earlier. The ten filters of the Vocoder synthesizer were widened to cover the range 0-7500 Hz, and a set of manual controls was supplied. The output amplitudes of the filters were controlled from a keyboard; a wrist bar selected buzz or hiss excitation; and a foot-pedal controlled F_0. There were also special keys to generate automatically the sequence of closure and release required for stops. A year or more of training was required before an operator could produce intelligible speech, and each utterance had to be carefully rehearsed. The Voder was demonstrated successfully at the 1939 New York World's Fair and the 1940 San Francisco World's Fair.

There were two important differences between Dudley's approach to speech synthesis and von Kempelen's. First, Dudley's Voder was an electrical, rather than a mechanical, simulation, so that the acoustic properties of synthesizer components were reasonably predictable and design changes could be readily made. Second, the Voder simulated acoustic properties of speech, whereas von Kempelen's simulated articulatory properties as well. Dudley's model made it easier to improve the rendition of particular speech sounds but made a weaker claim about the nature of speech. However, Dudley's system had one important feature in common with von Kempelen's: a human operator was used, and the rules for synthesis were part of his skill.

The human operator disappeared from speech synthesis as an indirect result of Potter's invention of the sound spectrograph during World War II (Koenig et al., 1946). After the War, the spectrograph opened the way to extensive research in acoustic phonetics because it made it easy to observe the correspondence between speech sounds and events in the acoustic spectrum, notably formant movements. The spectrograph also suggested a new way of synthesizing speech: "playing back" a spectrogram. Potter himself built a playback synthesizer (Young, 1948); Cooper (1950) developed a research version, the Pattern Playback, which is still in use at Haskins Laboratories. In the Pattern Playback, an optical representation of an excitation spectrum with 50 harmonics and F_0 at 120 Hz is shaped by a spectrographic pattern painted on a moving, transparent acetate belt,and this optical representation is then converted to an acoustic signal. Thus, the synthesis of an utterance is not a transient performance but is controlled by a preplanned pattern and can be repeated. Moreover, the close correspondence between the output of the analyzing tool (the spectrograph) and the input to the synthesizing tool (the Playback) is convenient experimentally and of great value conceptually.

The Haskins investigators used the Playback to study the psychology of speech perception and to accumulate a body of knowledge about the "speech cues" (Liberman et al., 1967). Experienced users of the Playback, the late Pierre Delattre, for example, could readily paint intelligible utterances: like the operators of von Kempelen's synthesizer or the Voder, they had internalized a set of rules. Frances Ingemann, however, used this body of knowledge to draw up a formal set of instructions for painting spectrograms (Ingemann, 1957; Liberman et al. 1959). Her instructions are subdivided into rules for manner class, place of articulation, and the voiced/voiceless distinction. (Since the rules are intended for a monotone synthesizer, there are no F_0 rules.) The manner-class rules specify steady-state durations of the associated formant transitions, and the formant amplitudes appropriate for each class. The place rules specify the steady-state frequencies of formants, fricative noise, and stop-bursts and the transition end-points for stops, nasals, fricatives, and affricates at each point of articulation. The voicing rules specify the burst durations and closure voicing for stops and the friction duration and intensity for fricatives. Because they are organized along the phonetic dimensions of manner, voice, and place, the rules make the most of the uniformities and symmetries which emerged from research on the speech cues. But this kind of organization could not be carried through consistently: separate rules were needed for the steady states of each vowel, and "position modifiers"--changes to the basic rules--were required in some contexts. These modifiers reflect basic limitations of synthesis by rule at the acoustic level.

Using these rules, the utterance "I painted this by rule without looking at a spectrogram and without correcting by ear. Can you understand it?" was synthesized on the Playback. Painting an utterance of any length to precise specifications, however, proved to be a laborious procedure, and not many other such utterances were actually synthesized. Nevertheless, for the first time, a set of rules had been stated explicitly enough so that anyone willing to take the trouble could paint a spectrogram by rule and synthesize the corresponding utterance, any spectrogram purporting to follow the rules could be checked, and utterances representing different versions of a rule could be directly compared. The concept of speech synthesis by rule, which had been the implied purpose of earlier investigators, now became a clearly understood research objective.

Meanwhile, the resonance synthesizer had been developed. This type of synthesizer derives from the formant vocoder just as the Voder derives from the filter-bank vocoder. Resonant circuits represent the first few formants, and these circuits are excited by a hiss source or a variable-frequency buzz source. The state of the synthesizer can thus be specified by assigning values to a relatively small number of parameters which correspond to significant dimensions of natural speech and are readily observable in acoustic records: F_0, the formant frequencies F_1, F_2, F_3, and so on. A much stronger claim is implicit in a parametric synthesizer than in a nonparametric synthesizer like the Voder or the Playback. The information of interest in speech is regarded not just as band-limited, but as entirely a function of a very few physical variables.[2]

The resonant circuits of the synthesizer can be arranged either in parallel, the outputs of the various circuits being summed to produce the final output, or in series, the output of each resonant circuit being fed into the next. The series synthesizer is a closer approximation to the acoustical behavior of the vocal tract and incorporates in its design Fant's (1956) observation that if certain assumptions are made about the glottal source spectrum and the formant bandwidths, the relative amplitudes of vowel formants can be predicted from their frequencies. Thus, a series synthesizer requires fewer parameters and can be expected to produce more natural vowels. On the other hand, parallel synthesizers are far more flexible and simplify synthesis strategy for sounds with complex spectra, like voiced fricatives. The relative merits of parallel and series synthesizers are best summed up by Flanagan (1957). Lawrence's (1953) PAT was the first example of a parallel resonance synthesizer; Fant's (1958) OVE II, the first full-scale, experimental series synthesizer. Highly reliable resonance synthesizers of both types are now available.[3]

---

[2] For an interesting discussion of parametric vs. nonparametric synthesis, see Ladefoged (1964).

[3] Other parallel resonance synthesizers are described by Borst, (1956); Holmes et al., (1964); Mattingly, (1968b), and Glace, (1968). Other series synthesizers are described by Coker, (1965); Tomlinson, (1965); Liljencrants, (1968); Kacprowski and Mikiel, (1968); Kato et al., (1968); Dixon and Maxey, (1970); Shoup, (pers. comm.)

A parametric control scheme should have made synthesis by rule simpler, since the parameters to be specified are precisely the dimensions of speech in terms of which it is convenient to state acoustic rules. Ingemann (1960), in fact, reformulated her rules for use with the Edinburgh-series version of PAT (Anthony and Lawrence, 1962). But in order to control a resonance synthesizer, some means of changing the parameter values dynamically is required. At first, this was accomplished with a function generator; for example, parameter functions for the Edinburgh PAT were represented in conductive ink on parallel tracks of a moving plastic belt (Fourcin, 1960). But *applying* the rules (as distinct from *stating* them) was, if anything, more troublesome with a function generator than with the Pattern Playback. Fortunately, digital computers now began to become available for phonetic research. Kelly and Gerstman (1961) demonstrated that the computer not only could apply a set of rules (i.e., calculate the parameter values) quickly and accurately but also could be used to simulate the synthesizer itself.[4] Other investigators showed that, if an actual, rather than a simulated, synthesizer is used, the computer could also play the role of function generator.[5]

The Kelly and Gerstman program was quite simple. For each speech sound, initial and final transition durations, steady-state durations, and steady-state values for each parameter were stored. During the steady state of a sound, the stored values were used; during the final-initial transition period, parameter values changed smoothly from preceding to following steady state. It would be easy to criticize this scheme: the framework within which the rules are stated is extremely crude, and a good deal of *ad hoc* modification was required to make the synthetic speech even reasonably intelligible. But Kelly and Gerstman had clearly demonstrated that a computer could be used to apply phonetic rules--as great an advance over application of the rules by drawing patterns or functions by hand as were the latter over direct operation of the synthesizer by a human being. It was now possible to test and correct rules by producing substantial quantities of synthetic speech automatically and consistently.

---

[4] The advantage of a simulation is that it can be completely reliable and accurate, and the design of the synthesizer can be readily modified; the disadvantage is that an extremely powerful computer is required, and such computers are too expensive to permit extended real-time operation. Recent simulations of resonance synthesizers (all series) include those described by Flanagan et al. (1962), Rao and Thosar (1967), Rabiner (1968), Saito and Hashimoto (1968).

[5] On-line transmission of stored parameter values can be performed by a laboratory computer at a low enough cost to permit the investigator to experiment at length; it is easy to program other convenient facilities, such as routines for editing or displaying the stored parameter values. Schemes of this sort include those of Tomlinson (1965), Denes (1965), Coker and Cummiskey (1965), Scott et al. (1966), Mattingly (1968b). Off-line control schemes, in which the computer produces a record, such as a paper tape, which is then used to control a function generator, are also practical, though less convenient (Holmes et al., 1964; Iles, 1969).

Resonance synthesizers, as well as the Playback and the Voder, are "terminal analog" synthesizers: they simulate the acoustic output of the vocal tract but not the activity of the vocal tract itself. However, concurrently with resonance synthesizers, vocal-tract analog synthesizers were being developed. With one interesting exception, the "true" (i.e., mechanical) model of Ladefoged and Anthony (Anthony, 1964), these synthesizers are electrical simulations. The supraglottal vocal tract is considered as segmented into a series of short tubes, each with a variable cross-sectional area. The acoustic properties of each tube in such a series can be simulated by the electrical properties of a transmission line. The acoustical effect of a change in cross-sectional area is equivalent to a change in the characteristic impedance of the corresponding transmission-line segment. The nasal cavity is usually represented as a branch with a few fixed sections and variable coupling to the main line. Thus, the spectrum of the output of the synthesizer depends on the momentary cross-sectional area function and the amount of nasal coupling.[6]

Like a resonance synthesizer, a vocal-tract analog could be simulated on a computer, and Kelly and Lochbaum (1962) used such a simulation for synthesis by rule. The approach was very much the same as the one used by Kelly and Gerstman, except that the parameters were the areas of the cross-sections of the segments of the tract instead of formant frequencies. The results were less successful than Kelly's terminal analog synthesis had been, a fact of some interest.

By the early sixties, then, there was no doubt that speech could be synthesized by rule by either terminal analog or vocal-tract analog methods. Reliable synthesizers and convenient methods of controlling them had been developed. Even more important, the value of explicitly formulated rules had become obvious.

JUSTIFICATION FOR SYNTHESIS BY RULE

Since speech has been successfully synthesized by rule, it might seem that the basic objective of von Kempelen and his successors has been attained; it therefore becomes important to state clearly the reasons for continuing the research. One obvious reason is that we still have much to learn about the physical aspects of speech production: how the various articulators move; how their movements are timed; how the controlling musculature operates to produce the sounds of speech. Synthesis by rule is a way of testing our understanding of the physical apparatus, and this is the primary motivation for much of the activity in the field today. But this argument may not seem very persuasive to the linguist, who is concerned with speech as a psychological fact rather than a physical one. But we think that synthesis by rule offers other possibilities of substantial theoretical importance for the linguist, possibilities which have barely begun to be explored. To justify this point of view, however, requires brief reference to some basic questions of linguistics.

---

[6] The first electrical vocal-tract analogs were static, like those of Dunn (1950), Stevens et al. (1953), Fant (1960). Rosen (1958) built a dynamic vocal tract (DAVO), which Dennis (1963) later attempted to control by computer. Dennis et al. (1964), Hiki et al. (1968), and Baxter and Strong (1969) have also described hardware vocal-tract analogs. Kelly and Lochbaum (1962) made the first computer simulation; later digital computer simulations have been made, e.g., by Nakata and Mitsuoka (1965); Matsui (1968), and Mermelstein (in press). Honda et al. (1968) have made an analog computer simulation.

We believe, following Chomsky and Halle (Chomsky, 1965, 1968; Chomsky and Halle, 1968) and other generative grammarians, that the grammar of a language can be represented by a set of rules. A speaker/hearer who is competent in a language has learned these rules and uses them to determine the grammatical structure of his utterances or those of another speaker, since the rules "generate" an utterance if, and only if, grammatical structure can be assigned to it. Competence does not fully determine performance: the speaker's actual utterances may frequently be ungrammatical, and the listener may guess a speaker's intent without consistent reference to the rules.

A subset of the rules for any language are phonological: they convert a string of morphemes, already arranged in some order by syntactic rules, into a phonetic representation. In the familiar generative model (Chomsky and Halle, 1968), the morphemes are lexically represented as distinctive-feature matrices, each column of which is a phonological segment. All phonologically redundant feature specification is omitted, and each specified feature has one of two values. The phonological rules complete the matrices, alter the feature specification in certain contexts, delete and insert segments, and assign a range of numerical values to the features. The output of the phonological rules, then, is a matrix of phonetic segments for which each of the features is numerically specified.

Besides these acquired rules, we suppose the speaker of the language to have a certain inborn linguistic capacity: "the innate organization that determines what counts as linguistic experience and what knowledge of language arises on the basis of this experience" (Chomsky, 1968:24). This capacity is what makes it possible for him to learn the rules and to use them in making grammatical judgments; it is reflected in some universal and quite severe constraints on the form and content of grammatical rules. In the case of phonology, not only are the forms of the input and output highly determined, but the set of phonetic features by which the output of the phonology may be represented is the same for all languages; moreover, the language-specific set of "classificatory" features which are used to represent the lexical items at the input to the phonology are related in a significant, though complex, way to a subset of universal phonetic features with the same names. Phonological rules are constructed out of phonetic raw material.

Conventionally, the linguist's concern ends with the phonetic-feature representation, which is the output of the phonology. But our account of the speaker/hearer's inborn capacity is incomplete, for we have said nothing about his knowledge (quite unconscious, but no less psychologically real) of the relationship between the phonetic-feature representation and the acoustic signal. The speaker/hearer can produce an acoustic representation of an utterance, given the feature representation (speech production), and he can apparently recover the feature representation for an utterance produced by someone else (speech perception). Neither process is trivial; to be able to produce and recognize speech, he must possess a definition of each feature in sufficient detail to enable him to assign a possible value to it, given the acoustic signal. He must, therefore, have sufficient information about the anatomy, physiology, and acoustics of the vocal tract to permit such a definition, and he must also understand how the apparently discrete representation of an utterance, at the output of the phonology, as a series of phonetic segments, is translated into a continuous representation in the acoustic signal. Perhaps it would not be inappropriate to picture the speaker/hearer's knowledge as consisting of a dynamic neural simulation of the vocal tract, the state of which is determined by the values of the features and which guides his production and perception of speech (Mattingly and Liberman, 1969). If we assume that both the features and the general structure of the human vocal tract are universal, it seems highly likely *a priori* that this knowledge of the speaker/hearer's is inborn; moreover, some experimental evidence for such a view has recently appeared (Moffitt, 1969; Eimas et al., 1970).

Just as we characterize phonological and syntactic capacity by assigning to them formal and substantive properties, in the form of conventions and features, so to the extent that we can describe this simulated vocal tract, we characterize what may be called "phonetic capacity." It is to phonetic capacity which Chomsky and Halle (1968:294-295) allude when they observe:

> The total set of features is identical with the set of phonetic properties that can in principle be controlled in speech: they represent the phonetic capabilities of man and, we would assume, are therefore the same for all languages.

We view the development of an adequate account of phonetic capacity as the chief goal of experimental phonetics. There appear to be two important tasks. First, it is necessary to determine the membership of the set of universal phonetic features, since these are the basis of phonological capacity and the elements of phonological competence. Moreover, we want to understand the role of the various stages in the speech chain in the psychological definition of each feature. These stages include (at least) the activation of the muscles of the vocal tract by neuromotor commands, the gestures made in response to these commands by the various articulators, the resulting dynamic changes not only in the shape of the vocal tract but also in air pressure and airflow at different points in the tract, and finally, the cues in the acoustic output. It may well be that not all these stages are pertinent to phonetic capacity; on the other hand, other stages, as yet poorly understood, may be involved.

The second task is to characterize psychologically the translation from the discrete, essentially timeless phonetic level to the continuous, time-bound activity characteristic of lower levels. If, at any of these levels, speech could be consistently separated into stretches corresponding to phonetic segments, the problem would be fairly simple, but we know that this cannot be done. At any level in speech that we can observe, there are no true boundaries corresponding to any phonetic unit shorter than the breath-group, though in the acoustic signal there are many apparent boundaries reflecting articulatory events, e.g., stop closures and releases, spectral discontinuities in liquid and nasal sounds, onset and offset of voicing, and the like. Yet the psychological reality of phonetic segments can hardly be doubted, and at any rate, without them, phonology would collapse.

The feature-specified, dynamic vocal-tract model by which we would represent phonetic capacity is on just the same level of theoretical explanation as the phonological and syntactic models of linguistics. It does not have, as yet, any but the most general sort of neurophysiological basis; it does not account in itself for productive or perceptual performance; in particular, it does not conflict with an analysis-by-synthesis account of speech perception. It is simply a way of stating some properties which the neural mechanisms for speech must incorporate to account for the observed behavior of speaker/hearers.

By virtue of his phonetic capacity, the speaker/hearer acquires certain skills, just as he acquires the phonological rules of his language by virtue of his phonological capacity. He must "calibrate" his perceptions to allow for the idiosyncrasies of the vocal tracts of the other speakers to whom he listens, even before he understands his native language. [If there were no other reason for postulating phonetic capacity, we would want to do so to account for the fact that an infant learns to interpret the output of the vocal tract of each of the individuals around him, though these vocal tracts differ radically from one another and from his own in size and shape; and he does so with sufficient accuracy to permit the collection of the "primary linguistic data" (Chomsky, 1965:25) essential for language acquisition.] Moreover, if he himself is to produce acceptable versions of the speech sounds he perceives, he also has to learn the idiosyncrasies of his own vocal tract. He must learn to control certain stylistic factors--speaking rate, attitudinal intonation, and so on--both in production and perception. Finally, he may need to learn certain global phonetic properties of his language, e.g., its "articulation basis" (Heffner, 1950:98-99).

In brief, we distinguish four distinct components underlying speech perception and production: 1) inborn phonological capability, 2) acquired phonological competence in one's language, 3) inborn phonetic capacity, 4) acquired phonetic skill.[7] For the study of these various components, speech synthesis by rule has certain impressive advantages.

First, we can hope to gain real understanding of the component of interest to us only by attempting a highly formal account; yet any nontrivial formal account will doubtless be quite complex: this is already apparent for phonological and phonetic capacities and particularly so for phonological competence--as a glance at the summary of rules in Chapter 5 of The Sound Pattern of English (Chomsky and Halle, 1968) will confirm--and must certainly prove true for phonetic skill as well. Much can be done to reduce apparent complexity by suitable notation. But, as in many other fields which make use of highly formal systems, checking the consistency of the formalization is most easily done by computer simulation. Linguists are, in fact, turning increasingly to computer simulation to check the operation of syntactic and phonological rules (Fromkin and Rice, 1970).

Second, various dependencies exist among the components. An account of the phonetic skill of the particular speaker must begin with some assumptions about his phonological competence in his language and his phonetic capacity. Only in terms of the former can idiolectal variations be defined; only in terms of the latter can speaking rate be discussed. Similarly, the rules by which we try to characterize phonological competence must be stated in a form determined by phonological capacity. Phonological capacity, finally, depends on the choice of a set of features, the interpretation of which is a matter of phonetic capacity. Given this kind of dependency, it seems extremely risky to try to form hypotheses about the nature of one component without being quite specific as to the assumptions being made about the others on which it depends. Yet this is an ever-present temptation. Speaker variation, for example, is investigated without specification of precise phonetic and phonological models. Structural linguists rightly incurred the censure of generative phonologists because they formulated their phonemic inventories without proper concern for phonological capacity; generative phonologists, in turn, might be criticized because the set of phonetic features, on which their much more principled account of phonological capacity depends, as yet lacks a fully satisfactory and explicit basis in phonetic capacity (Abramson and Lisker, in press). Obviously, it is very desirable to state clearly, when a certain component is being investigated, how this component is assumed to depend on other components.

---

[7] Tatham (1969a) has recently used the term "phonetic competence" to mean approximately what we mean by "phonetic capacity"; otherwise we might have used the former term rather than the asymmetrical "phonetic skill." Tatham's paper (see also Tatham, 1969b) contains some cogent arguments not only for the existence of phonetic capacity but also for its importance in the formulation of phonological rules in a natural way.

Third, the ultimate check of a hypothesis concerning any or all of the components is, of course, the intuition of the native speaker (Chomsky, 1965:21). However, the only reliable way to consult his intuition is to present him with speech which we have made sure conforms to our current phonetic or phonological hypothesis and find out whether he considers it well formed. To do this, however, we need carefully controlled speech stimuli (Lisker et al., 1962; Mattingly, in press).

Synthesis by rule is a technique which seems to meet these requirements. With the computer we can simulate our phonological and phonetic formulations rigorously; errors of form and logic come to light all too quickly. We are compelled to be explicit about the assumptions we make about other components; if they are simplistic or inadequate we will not be allowed to forget the fact. And we can check the native speaker's intuition directly by producing controlled synthetic speech.

Let us briefly consider what an ideal speech-synthesis-by-rule system would be like. It would, in the first place, simulate all the components we have just discussed. Phonetic capacity would be represented by a synthesizer and computer programs controlling it, capable of generating just those sounds which can be distinguished in production and perception by the speaker/hearer; phonological competence, by the rules of some language, stated in a form which would be an acceptable input to the system; phonological capacity, by a part of the computer program itself, which would impose severe limitations on the form or substance of the rules; and phonetic skill, by an additional set of rules specific to some particular speaker. The combined effect of all components should be such as to restrict the possible utterances to just those which are well-formed speech in a particular language (assuming appropriate syntactic and semantic constraints) from one particular speaker to another.

For each component, moreover, we would want to include all those aspects, and only those, which are relevant to the capacity and competence underlying his production and perception of speech. Suppose, for instance, (contrary to our present expectations) that, from a psychological standpoint, speech production proved to be only a matter of transmitting certain cues definable in acoustic terms and invariantly related to phonetic features and that speech perception consisted simply in detecting these cues. Our "neural vocal-tract simulation" could then be just a terminal analog synthesizer. There would then be no reason for including neuromotor commands, gestures, or shape change in a parsimonious synthesis-by-rule system, because these matters would be irrelevant to phonetic capacity. They might continue to be of great interest from the standpoint of the physiologist and acoustician interested in speech, but they would have no claim on the linguist's attention.

Our ideal system is not concerned with performance as such. Even though our model is dynamic and the output is audible, the process of synthesis is a derivation according to rules, not a life-like imitation of a speaker's actual speech behavior. The output is acceptable to the hearer because it follows the rules, not just because, on the one hand, it is intelligible, despite errors and deviations, or because, on the other, it is highly natural-sounding--though one might expect that the output of an ideal system would be natural-sounding, if not physically naturalistic. Here our emphasis differs somewhat from that of Ladefoged (1967) and Kim (1966) who share our conviction that it is important to do synthesis by rule, but for whom linguistic and phonetic theory "must lead to the specification of actual utterances by individual speakers of each language; this is physical phonetics " (Ladefoged, 1967:58). From our point of view, it is not physical realism but psychological acceptability that is the proper evidence for correctness at the phonological and phonetic levels, just as it is on the syntactic level.

In the preceding discussion, we have deliberately generalized the concept of "synthesis by rule" to embrace phonology and phonetics. It would be possible to generalize still further to include syntax and semantics in a synthesis-by-rule system. But while computer simulations of syntactic and semantic rules are certainly desirable, the motivation for coupling them to a phonological and phonetic synthesis-by-rule system is less compelling, primarily because a set of syntactic rules can, in practice, be evaluated more or less independently of the associated phonology and phonetics.

CURRENT WORK IN SYNTHESIS BY RULE

We turn now to an assessment of the progress that has been made toward the ideal which has just been sketched. The first thing to be said is that most of the activity and most of the progress so far falls under the heading of phonetic capacity. Since the other components all depend, directly or indirectly, on phonetic capacity, this is just as it should be. Moreover, since we want to assess the role of the different stages of the speech chain in phonetic capacity, it is good that, in the present state of our knowledge, the research has been pluralistic: different types of systems have been developed in which the contribution of different stages has been emphasized. This has been difficult to do because appropriate data on which to base investigations at stages before the acoustic stage are hard to collect. At present, most of the work has been at the acoustic stage; the relationship between shape and acoustic output is quite well understood and several synthesis-by-rule systems operating on vocal-tract shape have been developed; systems which represent the movements of the actual articulators are beginning to show results; and some work has been done at the neuromotor command stage.[8]

Acoustic-Level Systems

We have already mentioned some systems in which the phonetic level is mapped directly onto the acoustic level, including one, that of Kelly and Gerstman (1961), which, like other more recent systems of this kind, is parametric. In these systems a target spectrum for each phone[9] is specified by a set of stored parameter values. Given a phonetic transcription of an utterance, the synthesis program calculated the momentary changes of value for each parameter from target to target as a function of time. (Notice that this is an extremely natural way to treat the problem of translating from the discrete to the continuous domain.)

The most important differences among the various systems have to do with the procedures for this calculation and, in particular, the procedure for calculating formant motion, since intelligibility depends crucially on the choice of targets toward which the formants move and the timing of their movements. In the Kelly-Gerstman system, it will be recalled, an initial transition duration, a final transition duration, and a steady-state duration are stored for each phone. The duration of a transition between two adjacent phones is the sum of the final transition duration of the first phone and the initial transition duration of the next. During the steady-state period, formants remain at their target values; during the transition period, they move from one set of target values to the next, following a convex path from consonant to vowel, a concave path from vowel to consonant, and a linear path otherwise.

In the system of Holmes et al. (1964), a "rank" is stored for each phone, corresponding to its manner class. Manner classes having characteristic transitions (e.g., stop consonants) rank high; manner classes for which the transition is characterized by the adjacent phone rank low. The character of the transition between adjacent phones is determined according to the ranking phone. Each transition is calculated by linear interpolation between a target value for the first phone and a boundary value and between the boundary value and the target value for the second phone. The durations of the two parts of the transition are stored for the ranking phone. The boundary value is equal to $C_R + W_R$ (F_A), where C_R is a constant and W_R a weighting factor for this formant stored for the ranking phone, while F_A is the target value of this formant stored for the adjacent phone. Hence, the character of the transition depends mainly on variables stored for the ranking phone. Thus, each phone has within its boundaries an initial transition, influenced by the previous phone, and a final transition, influenced by the following phone. A duration is stored for each phone; if it is greater than the sum of the durations of the initial and final transitions calculated for the phone, the target values are used for the steady-state portion. If the duration is less than the sum and the paths of the calculated transitions fail to intersect, they are replaced by a linear interpolation between the initial and final boundary values. But if the paths do intersect, the values for each transition between the boundary value and the intersection are used and the others discarded. Thus, the formants of shorter vowels do not attain their targets; their frequencies are context-dependent, as in natural speech (Shearme and Holmes, 1962; Lindblom, 1963).

---

[8] There is, of course, another way to synthesize speech by rule, and that is to compile an utterance from an inventory of shorter segments, themselves either natural or synthetic. Such approaches may have practical value, but from a theoretical standpoint they merely serve to remind us that there is no simple correspondence between phones and segments of the acoustic signal. See the discussion in Liberman et al. (1959). Systems in which speech is compiled from natural segments have been described by Harris (1953), Peterson et al. (1958), and Cooper et al. (1969). Systems using synthetic segments are described by Estes et al. (1964), Dixon and Maxey (1968), and Cooper et al. (1969).

[9] Workers in synthesis by rule (including the author) have been in the habit of referring to the units of their input transcriptions as "phonemes." In most cases, these units do not correspond either to the phonemes of structural linguistics or to the phonological segments of generative phonology; they tend to be closer to the level of a broad phonetic transcription. We use the term "phone," except in the case of systems, where a deliberate distinction is attempted between phonological and phonetic levels.

Denes (1970) uses a similar scheme, the boundary values being dependent on the target values and on a weight assigned to each phone. Our own system (Mattingly, 1968a,b) also uses a scheme like that of Holmes et al., except that interpolation is done according to a simple nonlinear equation which assures that formants curve sharply near boundaries. The formant transitions in Rabiner's (1967) system, the most serious attempt to simulate natural formant motion, are calculated according to a critically damped, second-degree differential equation. The manner in which a formant moves from its initial position toward the next target depends on a time constant of the equation, which is specified for each formant and each possible pair of adjacent phones. When all formants have arrived within a certain distance of the current target, they start to move toward the following target, unless a delay (permitting closer approximation or attainment of the target) is specified. It is not obvious that schemes for nonlinear motion offer any great advantage over linear schemes. While a nonlinear rule results in formant movements which are more naturalistic, they do not seem to be necessarily perceptually superior to, or even distinguishable from, linear movements. If the formant moves between appropriate frequencies over an appropriate time period, the manner of its motion does not seem to be too important.

In Rao and Thosar's (1967) system, each phone is characterized by a set of "attributes," i.e., features of a sort. A phone is either a vowel or a consonant; vowels are front or back; consonants are stops or fricatives, voiced or unvoiced, labial, dental, or palatal. Transition patterns depend on these attributes and on the duration and steady-state spectral values stored for each phone. Vowel-vowel transitions are linear from steady state to steady state, and the two temporal variables--total transition time and the fraction of the total within the duration of the earlier vowel--are the same for all pairs of vowels. For consonant-vowel transition, the boundary value for each formant is equal to $F(F_L) + (1-F)F_V$, where F_V is the target frequency of the vowel, F_L is the consonant locus frequency, and F is a weighting factor. Transition time and F_1 locus depend on the value of the stop-fricative attribute; F_2 and F_3 loci and the weighting factor, on the place-of-articulation attribute. Given the boundary value, steady-state values, and transition times, transitions are calculated as by Holmes et al.

Rao and Thosar resort to stored values for vowel spectra and vowel durations; Kim (1966), however, proposes that even these matters can be systematically treated. For example, his translation from distinctive-feature values to formant frequencies is made by defining the features in terms of "degrees" of difference from the [ə] frequencies. From the value assigned to one degree, and the [ə] frequencies, the frequencies of other vowels are calculated by means of such rules as "if High, -2d." The formant-frequency values determined in this way agree well with the data in the literature. However, since the degree values are not predicted on any principled basis, but are arrived at inductively by an averaging procedure applied to this same data, the agreement is hardly surprising and does not represent any interesting advance over stored values.

Several of these systems have been empirically successful in that they have proved capable of consistently producing intelligible speech. They also have enough theoretical plausibility to be used in investigations of other components. One could, for example, use them to test phonological rules proposed for a language (Mattingly, in press). But they are still inadequate because their working assumption is that phonetic capacity can be adequately described at the acoustic level. If this were so, a simple and consistent correspondence would hold between phonetic features and acoustic events. But in fact the correspondence is only partial. On the one hand, certain regularities are observable, which can be exploited in a synthesis-by-rule system, as Liberman et al. (1959) pointed out: F_1 and F_2 transitions and the type of acoustic activity during stop closure provide a basis for a purely acoustic classification of labial, dental, and velar voiced stops, voiceless stops, and nasals. On the other hand, the cues for a particular feature, regarded simply from an acoustic standpoint, are a rather arbitrary collection of events. There seems to be no special reason why a fall in F_1, a 60-150 msec gap, a burst, and a rise of F_1 should all be cues for a stop consonant, and no obvious connection between the locus frequency and the burst frequency of a stop at the same place of articulation. These cues only make sense in articulatory terms. Still, the apparent arbitrariness of the cues should not, in itself, discourage the formulation of acoustic rules for features. A more serious difficulty is that, in many cases, features cannot be independently defined at the acoustic level. Thus the voiced/voiceless distinction is cued in one way for stops and in another for fricatives. The frequencies at which noise is found in a fricative do not correspond to the frequencies of either the locus or the burst of a stop at a similar point of articulation. The frequencies of the first and second formants are sufficient to distinguish the nonretroflex vowels, but the range of F_1 variation seems to be influenced by the F_2 value: the vowels are not distributed regularly in F_1/F_2 space. Because of these difficulties, most of the acoustic synthesis-by-rule systems provide only for a regular relationship between phones and acoustic events; they do not attempt to define a set of acoustic features. The simple system of Rao and Thosar is exceptional in that (like Ingemann's set of Playback rules) it tries to make the most of the regularities which do exist, but the idiosyncratic characteristics of each phone must also be specified by these investigators.

The manner in which these systems translate from the discrete phonetic level to the continuous acoustic level also proves somewhat unsatisfying. The notions "target" and "transition" imply that the former characterizes essential aspects of a phone and the latter is a means of connecting one phone smoothly with another. In fact, as is well known, much of the information at the acoustic level in speech is encoded in the formant transitions, and most of the ingenuity devoted to acoustic rules has had the purpose of providing appropriate transitions for the various form and manner classes. This circumstance does not invalidate the notions "target" and "transition" for synthesis by rule in general; it is merely a further indication of the inadequacy of acoustic synthesis by rule.

Vocal-Tract Shape Systems

It appears, then, that there are limitations on the adequacy of a synthesis-by-rule system operating only with the acoustic stage. A number of systems have therefore been developed which incorporate earlier stages in the speech chain.

The next earlier stage in the speech chain is vocal-tract shape, which, for a given source of excitation, determines the spectrum of the acoustic output (Fant, 1960). Since the acoustics of speech production is complex, it seems plausible that rules for synthesis could be more readily and simply stated in terms of dynamic variations in shape. The speech implied by a sequence of shapes can then be heard with a vocal-tract analog synthesizer.

The general strategy used for synthesis by rule with a vocal-tract analog, which parallels the strategy used for acoustic systems, has been to specify a target shape for each phone and to interpolate by some rule between targets. In the system of Kelly and Lochbaum (1962), transition times and target shapes, represented as area functions, are stored for each phone. During the transition, the series of area values for each segment of the vocal-tract analog (and also values for excitation parameters and nasal coupling) are obtained by linear interpolation between the target values. There are numerous exceptions to this general principle of operation, most of which are attempts to provide for the effects of coarticulation and centralization. Vowels next to nasal consonants are nasalized throughout. Labials do not have a fixed target shape: the lips are constricted or closed for a period during which the rest of the tract moves from the previous to the following target. An unstressed vowel has zero duration, and its target shape is the average of the shape for the corresponding stressed vowel and that for the neutral or [ə] vocal-tract shape. Separate target shapes are provided for velars before front, central, and back vowels.

Mermelstein's (in press) system follows a similar plan. Two lists serve as input to this system. The first is a table of the area-function values for an inventory of shapes; the second includes a series of target shapes, specified with reference to the first list and corresponding to phones or temporal segments of phones, target values for other parameters, and transition durations. Mermelstein uses linear transitions near sharp constrictions and exponential transitions during periods when the shape of the tract is changing more slowly. He finds that this procedure, which effectively avoids steady states, contributes considerably to the naturalness of the speech.

Nakata and Mitsuoka (1965) use a more elaborate transition procedure based on a conception of Öhman (1967). In the case of vowel-to-vowel transitions over a period t', the momentary area function for a vowel at $\propto$t,

$$^{V}A(t') = {^{V}A^{T}} + ({^{V}A^{O}} - {^{V}A^{T}})\, W_K(\propto t')$$

where $^{V}A^{O}$ is the starting value, $^{V}A^{T}$ is the target value, and $W_K(\propto t')$ an asymptotic weighting function equal to 1 at starting and 0 at target. In the case of a consonant between two vowels, the effect of superposition of the consonant is taken as equivalent to the effect over a period θ of the consonant on the neutral tract,

$$CA(\theta) = {^{V}A_N} + [{^{C}A} - {^{V}A_N}]\, W_C(\theta)$$

where ${}^{V}A_N$ is the neutral tract, ${}^{C}A$ the consonant configuration, and $W_C(\Theta)$ another weighting factor. The result of superposition

$$\begin{aligned} CA(t') &= {}^{V}A(t') + {}^{C}A(\Theta) - {}^{V}A_N \\ &= {}^{V}A(t') + [{}^{C}A - {}^{V}A_N]\, W_C(\Theta) \end{aligned}$$

[Ichikawa and Nakata in a later paper (1968) treat superposition as multiplicative rather than additive.] Nakata and Mitsuoka claim that this rule automatically gives a good approximation of the different shapes of [k] in [ki] and [ko] at the time of maximal constriction, a fact which acoustic systems and earlier articulatory systems handle ad hoc.

The obvious advantage of using shape rules rather than acoustic rules is that the translation from the discrete to the continuous domain becomes more straightforward. The rule for transitions for stops can be stated simply and in the same terms as the rules for glides. But the notion of a target shape is rather unsatisfactory, because only a certain part of the shape is pertinent to any particular phone, and the rest must be arbitrarily specified. In another sense, moreover, a shape model is less interesting than an acoustic system. Ladefoged (1964) has pointed out that synthesis systems may be classified both as articulatory or acoustic and as parametric or nonparametric. The shape models we have just been discussing are articulatory, but they are not based on any natural parameters comparable to formant frequencies, still less on any set of features. Instead, an arbitrary number of vocal-tract cross-sections is used. This is a level of development corresponding to the point in acoustic phonetics when the most significant possible representation of the acoustic spectrum was in terms of a bank of filters. A further limitation of shape systems is that the transitional rules can be little more than arbitrary smoothing rules; it would be very difficult to characterize the changes in shape of the vocal tract differentially segment by segment. In fact, it is a question whether vocal-tract shape, as such, is a significant stage in the speech chain, except in a strictly physical sense. What is needed is a set of parameters for vocal-tract shape which would account for the behavior of the tract in the formation of the various sounds and at the same time facilitate a simple statement [illegible] rules.

Stevens and House (1955) have suggested a simple three-parameter model for vowel articulation in which the vocal tract is idealized as a tube of varying radius. Two of the parameters are $\underline{d}$, the distance of the main constriction from the glottis, and $\underline{r}_0$, the radius of the tube at this constriction. The radius $\underline{r}$ at another point along the tube depends on $\underline{r}_0$ and the distance $\underline{x}$ from the constriction: $\underline{r} - \underline{r}_0 = .25(1.2 - \underline{r}_0)\ \underline{x}^2$. The front portion of the tract (14.5 cm from the glottis and beyond), however, is characterized by a third parameter $\underline{A}/\underline{l}$, the ratio of the area of mouth opening to the length of this position of the tract. This ratio, inversely proportionate to acoustic impedance, varies depending on the protrusion of the lips. These parameters correspond, of course, to the familiar phonetic dimensions of front/back, open/close, and rounded/unrounded and serve to characterize vowels very well. With a static vocal-tract analog, Stevens and House were able to use this model to synthesize vowels with the formant-frequency ranges observed by Peterson and Barney (1952). These parameters are not, of course, satisfactory for most consonants, if only because the formula for computing $\underline{r}$ would break down under the circumstances of fricative narrowing and stop closure. Ichikawa et al. (1967) propose another, more general scheme for which the parameters are the maximal constriction point P and the maximum area points V_1 and V_2 of the front and back cavities formed by this constriction. Ichikawa and Nakata (1968) report that they have used this very over-simplified model in a synthesis-by-rule system. It does not seem likely, however, that any parametric description of vocal-tract shape will prove satisfactory unless it directly reflects the behavior of the various articulators in some detail. As Ladefoged (1964:208) has observed, "describing articulations in terms of the highest point of the tongue or the point of maximum constriction of the vocal tract is rather like describing different ways of walking in terms of movements of the big toe or ankle." But this is as much as to say that it is necessary to go back to the next earlier stage of the speech chain, the stage of articulatory gesture.

Articulator Systems

A number of investigators are attempting to write rules for synthesis in terms of the movements of the individual articulators. The basic approach is to assume a model for the motion of each articulator that is convenient for the statement of the rules. From the states of the articulator models, the vocal-tract shape and, in turn, the acoustic signal can be determined for a given excitation.

Coker (1967, pers. comm.) uses a modified version of a model suggested by Coker and Fujimura (1966). Two parameters for the lips, one for the velum, and four for the tongue determine the shape of the oral tract. The lip parameters indicate the degree of protrusion and of closure; the parameter for the velum indicates its relative elevation; two of the tongue parameters indicate the degree of apical closure and front-back position for the tongue tip; and the other two, the position of the central mass of the tongue in the midsagittal plane. For each phone, target values for these parameters are stored. The stored values are divided into "important" and "unimportant"; thus, degree of rounding is unimportant for most sounds but important for [i] and [w] at one extreme and [y] at the other. Interpolation from target to target is accomplished by a "low-pass filter" rule, which produces a certain amount of coarticulation and vowel reduction. The different parameters move at different speeds--for example, the apical parameter is quite fast and the protrusion parameter quite slow. The degree of coarticulation is greater for slowly moving parameters than for fast ones. Parameter speed is increased in transitions from unimportant to important values and reduced in transitions from important to unimportant values, thus increasing coarticulation for those parameters which specially characterize a particular phone. Parameter timing can also be modified depending on context; this feature of the system is used to provide anticipatory rounding. Target values for each phone are changed simultaneously, except that, in a consonant cluster, parameters for different articulators overlap. For each momentary set of articulatory parameter values, the corresponding vocal-tract shape is determined, and from the shape, the formant frequencies, which are used to control a resonance synthesizer.

Haggard (Werner and Haggard, 1969) has developed a similar model with 11 parameters. Like Coker, he has parameters for lip protrusion and lip closure, for elevation of the velum, and for tongue-tip position and closure. Position, degree of closure, and length of closure are parameters for the body of the tongue, and degree of closure for jaw and glottis. From a momentary description in terms of articulatory parameters, "constriction" (i.e., shape) parameters are derived which describe the vocal tract as a sequence of a few tubes of varying length and cross-sectional area. A nomogram of the sort given by Fant (1960:65) is used to calculate formant-frequency values for control of a resonance synthesizer. Target values of articulatory parameters stored for each phone are distinguished as "marked" or "unmarked," depending on whether they are characteristic of the articulation of the phone; the distinction is much the same as Coker's important vs. unimportant. A further distinction is made between position and closure parameters. The transition of a parameter from the midpoint of one phone to the midpoint of the next is made up of linear segments and varies depending on the type of each phone (vowel, consonant, or pause), the marking of the two target values, the characteristic rate of each parameter, and the parameter type (position or closure). The rather complex transition rules insure that marked target values for consonant closure parameters will be attained and held and that progress toward other marked target values will occur over a longer time and at a more rapid rate than toward unmarked values. Thus coarticulation and centralization are provided for. In general a phone can influence only the adjacent phones, but nasalization is provided for by allowing the velar closure parameter to influence several preceding phones.

Henke's (1967) model attempts to handle the same coarticulatory phenomena as Haggard's and Coker's, while avoiding a commitment to a parametrization in favor of a naturalistic representation of articulation. Each articulator is represented by a family of "fleshpoints" on the midsagittal plane. During the motion of an articulator, each point moves along a vector determined by a target location and a target articulator shape. During the early part of its motion, a point first accelerates as the inertia of the articulator is overcome, then attains an appropriate steady velocity, and finally slows as it approaches the target point. The motion of the articulators is determined by a set of attributes stored for each phone. A configurative attribute corresponds to a target location and shape; a strength attribute, to the force which moves an articulator. At any moment, motion may be controlled by attributes associated with one or several successive phones. However, different attributes referring to the same articulatory region cannot both apply at once. A change of attribute will occur at a time dependent on the attributes of the current phone and of the following phone and upon the progress of articulatory movements determined by other attributes. For example, when articulation of a stop consonant begins, the relevant stop attributes, specifying the shape and location of the articulator and the force of the closure, assume control of the articulator. When closure is attained, the attributes of a following vowel, except those which conflict with the stop attributes, are applied, and attributes of earlier phones are dropped. After the stop is released, all the attributes of the following vowel apply for enough time to allow the articulators to approach the vowel target.

Systems such as those of Coker, Haggard, and Henke are more theoretically adequate than shape systems; we are clearly closer to the level of phonetic features. The translation from discrete to continuous domains is more natural because a target is defined for each articulator. We might compare the kind of description given by these systems to that of an idealized X-ray movie of the vocal tract, from which not only dynamic changes in shape but also the contribution of each of the individual articulators to the changes in shape is apparent.

Neuromotor Command Synthesis

But the description is still deficient in some respects. The parameters that describe articulatory motion may seem the obvious ones and may be empirically successful, but they have no necessary theoretical basis. The various articulators, of course, are not free to move at random but only to and from a limited number of targets. This limitation on the number of targets accounts for the limited number of values that can be assumed even by features associated with such a complex articulator as the tongue. If we could go a stage further back in the speech chain and synthesize speech at the level of the neuromotor commands that control the muscles of the articulators, we might be able to account for these significant limitations. Fortunately, electromyographic techniques can help us here (e.g., Harris et al., 1965; Fromkin, 1966). From measurements of the voltages picked up during speech by electrodes placed in the vocal tract, it is possible to make some plausible inferences about muscle activity--and hence about the corresponding neuromotor commands--in the production of the sounds of speech.

The synthetic counterpart of electromyographic analysis would describe speech in terms of a series of commands to the muscles of the vocal tract. An approach to this kind of synthesis has been made by Hiki, who has developed a description of jaw and lip movement using muscle parameters (Hiki and Harshman, 1969). The forward part of the vocal tract is treated as an acoustic tube of varying length, height, and width. The value for each dimension depends on the positive or negative force exerted by lip and jaw muscles, and each of these muscles may affect other dimensions as well. Muscles of the lips that affect the same dimensions in the same way are grouped together, and the same is the case for muscles of the jaw. The force exerted by such a group of muscles (actually the effect of several neuromotor commands) is a parameter of the system. Four lip and two jaw parameters are used. The forces acting separately on lip and jaw are combined to produce a description of shape, and with this partial model, labial sounds can be synthesized with a vocal-tract analog. More recently Hiki has extended his investigations to the tongue (Hiki, 1970).

Clearly, synthesis by rule must move in the direction suggested by Hiki's work. Only with models of this sort, making use of the earliest observable stage of the speech chain, will it be possible to gain insight into the nature of individual gestures and their relative timing. It is significant, however, that myographic synthesis, as represented by Hiki's scheme, seems to lead to an increase rather than a decrease in the number of parameters, as compared with articulatory models, even though several muscles exerting parallel forces are grouped under one parameter. Though the neuromotor commands for lip closure, for example, are similar for the different manner classes of labial sounds (Harris et al., 1965), the relationship between this gesture and the neuromotor commands which produce it is not a simple one. This suggests that the connection between the phonetic feature corresponding to lip closure and the neuromotor commands may not be simple either; perhaps the realization of some value of a phonetic feature as a unitary psychological gesture may actually involve a complex neuromotor program. This view is reinforced by the recent finding of MacNeilage and DeClerk (1969) that coarticulation appears even in electromyographic data.

Synthesis of Excitational and Prosodic Features

Our discussion so far has been concerned with the synthesis of segmental phones and with supraglottal articulation and its acoustic consequences. A synthesis-by-rule scheme also has to take into account excitational, prosodic, and demarcative features, the associated glottal and subglottal events, and the acoustic correlates of these events.

Both resonance and vocal-tract analog synthesizers provide periodic and noisy excitation sources, periodic excitation being used for vowels, sonorants, and voiced stops; noisy excitation, for [h], aspiration, and frication. In resonance synthesizers, separate circuits (either fixed filters or variable-frequency resonators) are ordinarily provided for shaping high-frequency frication; in vocal-tract analog synthesizers, noise is inserted at various segments in the tract, depending on the place of articulation of the fricative. With such facilities the different kinds of excitation are readily simulated; the only problem is to write rules for the changes from one excitation source to another. This aspect of synthesis by rule has not been taken very seriously; usually the duration of the excitation appropriate for a phone is identical with the nominal duration of the phone itself. In the case of voiceless stops, however, this approach requires including part of the transition to the following vowel in the stop, as was done by Holmes et al. (1964). Another solution is to specify, as a characteristic of the voiceless consonant, the appropriate amount of devoicing of the following phone, as we have done (Mattingly, 1968a). What is really required, however, is a rule specifying voice-onset time negatively or positively relative to the instant of release, as the work of Lisker and Abramson (1967) suggests. For medial and final voiced consonants and consonant clusters, increased duration of the preceding vowel is well known to be an important cue (Kenyon, 1950:63; Denes, 1955), and some systems have taken account of it (e.g., Mattingly, 1968a; Rabiner, 1969).

Rather more attention has been given to prosodic and demarcative features such as stress, accent, intonation, juncture, and pause, which interact with inherent properties of a phone to determine duration, fundamental frequency, and intensity.

In our own prosodic control scheme for British English (Mattingly, 1966), two degrees of stress and three common intonation contours (fall, fall-rise, and rise) can be marked in the input. The F_0 rules specify a falling contour during the "head" of the breath group; the slope varies with the quality of the syllable nucleus. Voiceless consonants cause a "pitch skip"; stressed syllables, a smooth rise in F_0. The required terminal intonation contour is imposed on the "tail"--the last stressed syllable and any following syllables. Each possible syllable nucleus has an inherent duration which is increased multiplicatively by stress. In prepausal syllables, the duration of all phones is increased and amplitude is gradually diminished. More recently (Mattingly, 1968a), we have used similar rules for synthesis of General American and, in addition, provided for the durational effects of juncture.

Rabiner (1969) follows the model proposed by Lieberman (1967) in which the overall fundamental contour is determined by subglottal air pressure, except for the so-called "marked" breath group, where laryngeal tensing produces a terminal rising contour. Four degrees of stress can be indicated; the higher the stress, the greater the increase in F_0 on the stressed syllable. Duration increases additively with the openness and tenseness of the vowel and the degree of stress, as well as being affected by the following consonant.

Hiki and Oizumi (1967) have developed prosodic rules similar to those of Rabiner and Mattingly. The F_0 rules deal with pitch accent, emphasis, terminal contours, the overall contour, and individual differences; the duration rules take into account the inherent duration of phones, pause length and accent, and, interestingly, the effect of changes of tempo on these features.

Umeda et al. (1968) determine prosodic patterns for English directly from ordinary printed text and use them to control the vocal-tract synthesis-by-rule scheme of Matsui (1968). Syntactic rules of a primitive kind are used to divide an utterance into blocks corresponding to breath groups. An overall F_0 contour is imposed on each breath group. Word stress (along with the phonetic transcription for a word) is determined by table lookup, and the fundamental frequency and duration are increased accordingly. Intonation contours and pause duration are derived from the punctuation, if any, following each block.

Vanderslice (1968) has also considered prosodic features from the standpoint of the problems involved in the conversion of orthographic text to sound, correctly distinguishing some features not included in earlier systems and proposing rules for their synthesis. Two degrees of pitch prominence are used instead of one: "accent" and "emphasis," the latter for contrastive and emphatic stress. The features "cadence" and "endglide" replace the traditional fall, fall-rise, and rise, cadence being equivalent to a fall, endglide to a rise, and cadence followed by endglide to fall-rise. "Pause" is a separate feature. To account for the raising of F_0 in quoted material and its lowering in parenthetical material, the features "upshift" and "downshift," respectively, are used. An additional group of "indexical" features is proposed for stylistic variation, e.g., "dip," for the downward pitch prominence noted by Bolinger (1958).

While only a few terminal intonation contours carry grammatical information and are required for synthesis of ordinary discourse, many more occur in colloquial speech. Iles (1967), following a scheme of tones (i.e., terminal contours) proposed by Halliday (1963), has attempted to synthesize some of these tones, imposing the contours on segmental synthetic speech generated by PAT in the manner of Holmes et al. (1964).

The models of the behavior of F_0 discussed so far assume a basic contour for the whole breath group, on which stress and terminal intonation contours are imposed--an approach consistent both with the work of British students of intonation from Armstrong and Ward (1931) to O'Connor and Arnold (1961) and with the "archetypal" model of intonation proposed by Lieberman (1967). Another possible approach is to characterize a breath group as a series of pitch levels, as in Pike's (1945) well-known scheme. Shoup (pers. comm.) has synthesized sentences in which the F_0 contours corresponding to such descriptions are realized, taking into account the stress, the vowel quality, the excitation of the preceding consonant, and the frequency at the beginning of the syllable.

Synthesis by rule of prosodic features has come to receive serious attention only quite recently, by comparison with synthesis of segmental features. We have only just begun to understand what is easy and what is difficult, what is relevant and what is irrelevant. Of the three major correlates of the prosodic features, intensity has proved the least sensitive and the least important. F_0 has attracted the most interest: considerable success has been attained in producing convincing stress and terminal intonation contours by rule, and the articulatory mechanism has been simulated. Duration, however, remains a serious problem. No one has yet produced even an empirically successful set of duration rules, and it is far from clear what theoretically adequate rules would be like. Presumably there are some durational effects which are really automatic consequences of the articulation: formant transition durations surely fall into this category. A second group of effects are truly temporal but subject to phonological rule: vowel length, for example. Finally, there are effects which are, to some extent, under the conscious control of the speaker: speaking rate, for instance. All these different effects are superimposed in actual speech; sorting them out is a major task for synthesis by rule.

In the prosodic schemes we have just been describing, even those which model supraglottal shape or articulatory movement, the prosodic features are still being simulated purely acoustically. No attempt is made to model explicitly the articulatory mechanisms which are responsible for the variation in the acoustic correlates. Unfortunately, the prosodic articulatory mechanisms are much less well understood than those which underlie segmental features, which explains in part the lack of unanimity concerning the appropriate treatment of prosodic features at the phonological level.

Flanagan, however, has made impressive progress with his computer simulations of vocal-tract excitation (Flanagan and Landgraf, 1968; Flanagan and Cherry, 1969). Voicing is represented by the output of a system consisting of two masses, corresponding to the vocal cords, oscillating so as to vary the cross-sectional area of the passage between them, corresponding to the glottis. At one end of the passage is a source of air varying in pressure, representing the lungs; at the other end is a vocal-tract analog. In response to the subglottal air pressure, the displacement of each mass increases, as does the air flow through the glottis. But this increase in air flow results in an increase in negative Bernouilli pressure between the two masses, reducing the displacement, so that oscillation occurs. The frequency of the oscillation varies with the subglottal pressure, with the size of the two masses and their stiffness (vocal-cord tension), and with the acoustic impedance of the vocal-tract analog, which depends on the phone being synthesized. Thus the model allows simulation of the separate roles of lung pressure and cord tension in determining F_0 and takes account of interaction with the supraglottal tract. The same model serves to simulate frication, which occurs at a constriction in the supraglottal tract when the constriction is sufficiently narrow and the pressure behind sufficiently great. When both glottal and fricative excitation are present, as in a voiced fricative, the pattern of pitch-synchronous bursts in the noise is simulated in the model.

Synthesis Using Phonological Rules

As we have just seen, a substantial amount of research effort in speech synthesis by rule has been concerned with what we have called phonetic capacity. Other components--phonetic skill, phonological competence and capacity--have received relatively little attention. There are various reasons for this. The quality of speech synthesized by rule has only quite recently been good enough to serve as a vehicle for research in these other components; moreover, many of those engaged in synthesis by rule have been content to operate with a fairly rough and ready view of phonology, because they are more interested in the physical aspects of speech, either acoustic or articulatory, than with phonetic capacity as such, or its relationship to other components. A few years ago this might have mattered much less; but recent impressive developments in generative phonology make it important that synthesis by rule display greater sophistication in this area if linguists are to take it seriously.

A few scattered efforts have been made. In our own work (Mattingly, 1968a), we have drawn a distinction between the synthesis-by-rule program with the associated hardware, representing the universal aspects of speech (phonological and phonetic capacity) and the rules of a particular language or dialect (phonological competence) which were an input to the program. In practice this distinction is not made consistently: certain matters are handled in the rules which more properly belong in the program, and conversely.

The system also provides a kind of primitive phonological frame-work, in that it allows the statement of ordered, context-dependent allophone rules which modify the stored data for synthesis of a phone. The description of the contexts in which a rule can be altered are built up from a limited set of binary contextual features, e.g., "prevocalic," "postvocalic," "stressed": these contextual features are part of the program. Thus, it is claimed that the nature of phonological capacity is such that only a small fraction of the conceivable contexts in fact occur in the phonological rules of natural language--a claim with which Chomsky and Halle (1968:400-401) appear to be sympathetic. The program has been used to synthesize both the General American and the Southern British dialects of English (Haggard and Mattingly, 1968).

In this system, the prosodic rules are phonetic: phonologically predictable stress and intonation effects must be marked in the input. Vanderslice (1968), however, has proposed a set of rules for predicting the occurrence, in English, of the prosodic features for which his definitions have been given above, in particular, accent. His strategy is to assign provisional accents to all lexically stressed syllables and then to delete certain of these accents. For example his "rhythm rule" deletes the middle one of three consecutive accentable syllables in the same sense group. If a word such as "unknown" is assumed to have two accentable syllables in its lexical form, this rule accounts nicely for the shifting stress in such words. Other rules proposed by Vanderslice rely on syntactic or semantic conditions; these conditions will somehow have to be marked at the input to the phonology.

Finally, mention should be made of Allen's (1968) programming of the Chomsky and Halle (1968) rules for the assignment of accent. At this writing, so far as we know, no one has thus far attempted a program for synthesis of English based on the Chomsky-Halle rules for segmental phonology or even a computer simulation with phonetic-feature matrices as output. Fromkin and Rice (1970), however, have developed a program for which the input format follows closely that of the Chomsky-Halle phonological conventions and permits the testing of a set of phonological rules.

SUMMARY AND CONCLUSIONS

We must now try to sum up the current state of synthesis by rule and to indicate the directions the work may be expected to take in the future. As we have seen, it is possible to synthesize speech by rule which is not only intelligible but also reasonably acceptable to a native speaker. Moreover, the trend of research in the past few years has been toward the development of systems of increasing phonetic sophistication with correspondingly greater theoretical interest. In the synthesis of segmental sounds, the emphasis has shifted from acoustic models to vocal-tract shape models, and from shape models to articulator models; a similar trend is evident in prosodic synthesis. Though much of this work is motivated, in the first instance, by an interest in the physical aspects of speech production, it is clearly also leading toward an increased understanding of phonetic capacity.

In the future, it is desirable, first of all, that synthesis by rule become a tool for the study of phonological capacity and competence. In practice, this would mean the development, according to the principles of modern generative grammar, of phonological descriptions of languages, the outputs of which would be converted to speech by a synthesis system. Such enterprises would be valuable for two reasons: first, they would tend to correct the present, rather off-hand conceptions of phonology entertained by many of those who are now doing synthesis by rule; second, they would, on the other hand, compel the phonologist to relate his descriptive rules to some clearly defined concept of phonetic capacity and permit him to test utterances produced by his phonological rules against the intuitions of the native speaker. When the generative grammarian is doing syntax, it is quite natural for him to offer examples in a form such that any native speaker can determine their grammaticality; the grammarian should be able to operate on the same basis when he is doing phonology, and he can, if he uses synthesis by rule. It would also seem desirable to increase considerably the number of dialects and languages for which rules for synthesis have been written. Most of the work thus far has been done in English and Japanese; many other languages should be synthesized as part of a general effort to explore the different versions of phonological competence.

The use of synthesis by rule to study phonological competence and capacity will, of course, compel attention to the central problem of phonetic capacity, that of enumerating and defining the universal set of phonetic features and in the process giving increased psychological meaning to the notion "feature." This means, in practice, the development of systems in which phonetic feature matrices are the input to the part of the program that simulates phonetic capacity. Though various feature-like entities have played a part in several of the systems we have discussed, none of these systems really represents a consistent attempt to synthesize speech using what the phonologist would regard as phonetic features. Many problems must still be worked out. For example, a feature involving a particular articulator can be equated with a gesture of the articulator toward a particular target, but the synthesis-by-rule system must somehow define just what it is that accounts for the psychological unity of this gesture, regardless of the original position of the articulator. For manner features, the problem is still more acute: the system must explain how features such as "continuant noncontinuant" can be given a plausible unitary definition in terms of phonetic capacity, even though physically quite different articulations may be used for the production of the various stops and continuants. If the feature is defined in part by feedback of some kind, this must be part of the synthesis system.

Nor is the matter of the translation from the discrete to the continuous as yet handled really adequately by the systems we have discussed. Having recognized that the targets for each articulator must be separately described, we must now try to account in some principled way for the coordination of the movements of the various articulators toward their targets. Present systems, in which each phone is dealt with in turn and is affected by the preceding and following phone but no others (with the exception of arrangements in some systems to nasalize several preceding phones), are too restrictive to account for the fact that coarticulation may extend over several phones (Kozhevnikov and Chistovich, 1965). The assumption of these systems is that the changing of targets for the various articulators is synchronized phone by phone--an assumption which works empirically after a fashion but which masks the real problems of how and to what extent the movements of articulators are synchronized and how much account phonetic capacity must take of the synchronization process. It has frequently been suggested that the syllable, which certainly seems to have psychological reality--and therefore some role in phonetic capacity--is the unit of coarticulation. Clearly there is a need of a synthesis-by-rule system that explores this possibility.

Another area that needs a great deal of further attention is the nature of the demarcation between phonetic capacity and phonological competence. We want to reflect this separation as clearly as possible in a synthesis-by-rule system; unfortunately, it is not always possible to distinguish in particular cases between "intrinsic" allophones (belonging to phonetic capacity) and "extrinsic" allophones (belonging to phonological competence) (Wang and Fillmore, 1961). Moreover, Tatham (1969b) has argued that, since such an "intrinsic" difference as front vs. back [k] can be distinctive in some languages, we have to provide for countermanding, in special cases, of a normal rule of phonetic capacity by phonological competence. This is actually, as Tatham points out, a problem relating to "markedness," an issue involving the relationship between the two components which has concerned phonologists from Troubetzkoy (1939:79) and the Prague School to Chomsky and Halle (1968:402ff.).

Finally, we can also look forward to increasing our understanding of the elusive matter of phonetic skill through synthesis by rule. The systems we have discussed all assume an ideal or at least typical speaker with a consistent style; questions of phonetic skill are avoided. But given some reasonably satisfactory representation of other components, we can begin to derive auxiliary sets of rules representing phonetic skill and consisting of a series of modifications to the parts of the system representing phonological competence and phonetic capacity. Suppose, for instance, that we wish to investigate the productive and perceptual factors of speaker variation. These are matters, in part, of the physical characteristics of the speaker (and so will involve adjustments of the synthesizer itself) but also of phonetic skill. At present the preferred methods of study are subjective ratings of speakers and examination of spectrograms. But it should be possible, using synthesis by rule, to try to mimic speakers and to study listeners' perceptions of such mimicry under quite specific assumptions about the speaker's and the listeners' phonetic and phonological capacity and the rules of their language.

Questions such as these make it apparent that synthesis by rule forces attention to precisely those phonetic problems which are fundamental to phonology. We hope that some phonologists will be sufficiently intrigued to join in the search for the answers.

REFERENCES

Abramson, A. S. and L. Lisker. In press. Laryngeal behavior, the speech signal and phonological simplicity. Proceedings of the 10th International Congress of Linguists, Bucharest, 1967.

Allen, J. 1968. A study of the specification of prosodic features of speech from a grammatical analysis of printed text. Unpubl. Ph.D. thesis, M.I.T.

Anthony, J. 1964. True model of the vocal tract. JAcS. 36.1037.

Anthony, J. and W. Lawrence. 1962. A resonance analogue speech synthesizer. Proceedings of the Fourth International Congress on Acoustics, A. Kjerbya Nielsen, Ed., Paper G43. Copenhagen: Organization Committee, Fourth ICA.

Armstrong, L. C. and I. C. Ward. 1931. A handbook of English intonation. 2nd Ed. Cambridge: Hoffer.

Atkinson, R. C. and H. A. Wilson. 1968. Computer-assisted instruction. Science 162.73-77.

Baxter, B. and W. J. Strong. 1969. WINDBAG-a vocal-tract analog speech synthesizer. JAcS. 45.309.

Bolinger, D. 1958. A theory of pitch accent in English. Word 14.109-149.

Borst, J. M. 1956. Use of spectrograms for speech analysis and synthesis. J. Audio Eng. Soc. 4.14-23.

Chomsky, N. 1965. Aspects of the theory of syntax. Cambridge, Mass.: M.I.T. Press.

Chomsky, N. 1968. Language and mind. New York: Harcourt.

Chomsky, N. and M. Halle. 1968. The sound pattern of English. New York: Harper.

Coker, C. H. 1965. Real-time formant vocoder, using a filter bank, a general-purpose digital computer, and an analog synthesizer. JAcS. 38.940.

Coker, C. H. 1967. Synthesis by rule from articulatory parameters. Conference preprints, 1967 Conference on Speech Communication and Processing, 53-63. Bedford, Mass.: Air Force Cambridge Research Laboratories.

Coker, C. H. and P. Cummiskey. 1965. On-line computer control of a formant synthesizer. JAcS. 38.940.

Coker, C. H. and O. Fujimura. 1966. Model for specification of vocal-tract area function. JAcS. 40.1271.

Cooper, F. S. 1950. Spectrum analysis. JAcS. 22.761-762.

Cooper, F. S. 1962. Speech synthesizers. Proceedings of the Fourth International Congress of Phonetic Sciences, ed. by Antti Sovijatvi and Pentti Aalto. 3-13. The Hague: Mouton.

Cooper, F. S., J. H. Gaitenby, I. G. Mattingly, and N. Umeda. 1969. Reading aids for the blind: A special case of machine-to-man communication. IEEE Trans. Audio. 17.266-270.

Denes, P. B. 1955. Effect of duration on the perception of voicing. JAcS. 27.761-764.

Denes, P. B. 1965. "On-line" computing in speech research. JAcS. 38.934.

Denes, P. B. 1970. Some experiments with computer synthesized speech. Behav. Res. Meth. & Instru. 2.1-5.

Dennis, J. B. 1963. Computer control of an analog vocal tract. JAcS. 35.1115.

Dennis, J. B., E. C. Whitman, and R. S. Tomlinson. 1964. On the construction of a dynamic vocal-tract model. JAcS. 36.1038.

Dixon, N. R. and H. D. Maxey. 1968. Terminal analog synthesis of continuous speech using the diphone method of segment assembly. IEEE Trans. Audio. 16.40-50.

Dixon, N. R. and H. D. Maxey. 1970. Functional characteristics of an on-line, computer-controlled speech synthesizer. JAcS. 47.93.

Dudley, H. 1939. The vocoder. Bell Labs. Rec. 18.122-126.

Dudley, H., R. R. Riesz,and S. S. A. Watkins. 1939. A synthetic speaker. J. Franklin Inst. 227.739-764.

Dudley, H. and T. H. Tarnoczy. 1950. The speaking machine of Wolfgang von Kempelen. JAcS. 22.151-166.

Dunn, H. K. 1950. Calculation of vowel resonances, and an electrical vocal tract. JAcS. 22.740-753.

Eimas, P., E. R. Siqueland, P. Jusczyk, and J. Vigorito. 1970. Speech perception in early infancy. Paper presented to the Eastern Psychological Association, April 1970.

Estes, S. E., H. R. Kerby, H. D. Maxey, and R. M. Walker. 1964. Speech synthesis from stored data. IBM J. 8.2-12.

Fant, C. G. M. 1956. On the predictability of formant levels and spectrum envelopes from formant frequencies. For Roman Jakobson, ed. by Morris Halle et al. 109-120. The Hague: Mouton.

Fant, C. G. M. 1958. Modern instruments and methods for acoustic studies of speech. Proceedings of the Eighth International Congress of Linguists, ed. by Eva Sivertsen, 282-358. Oslo: Oslo Univ. Press.

Fant, C. G. M. 1960. Acoustic theory of speech production. The Hague: Mouton.

Fant, C. G. M., J. Martony, U. Rengman, and A. Risberg. 1963. OVE II synthesis strategy. Proceedings of the Speech Communications Seminar, Paper F5. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology.

Flanagan, J. L. 1957. Note on the design of "terminal" analog speech synthesizers. JAcS. 25.306-310.

Flanagan, J. L. 1965. Speech analysis synthesis and perception. Berlin: Springer.

Flanagan, J. L. and L. Cherry. 1969. Excitation of vocal-tract synthesizers. JAcS. 45.764-769.

Flanagan, J. L., C. H. Coker, and C. M. Bird. 1962. Computer simulation of a formant vocoder synthesizer. JAcS. 34.2003.

Flanagan, J. L. and L. L. Landgraf. 1968. Self-oscillating source for vocal-tract synthesizers. IEEE Trans. Audio. 16.57-64.

Fourcin, A. 1960. Potential dividing function generator for the control of speech synthesis. JAcS. 32.1501.

Fromkin, V. A. 1966. Neuromuscular specification of linguistic units. L & S. 9.170-199.

Fromkin, V. A. and D. L. Rice. 1970. An interactive phonological rule testing system. Working papers in phonetics 14, 8. Los Angeles: UCLA Phonetics Laboratory.

Gariel. 1879. Machine parlante de M. Faber. J. Physique Théorique et Appliquée. 8.274-275.

Glace, D. A. 1968. Parallel resonance synthesizer for speech research. JAcS. 44.391.

Haggard, M. P. and I. G. Mattingly. 1968. A simple program for synthesizing British English. IEEE Trans. Audio. 16.95-99.

Halliday, M. A. K. 1963. The tones of English. ArchL. 15.1-28.

Harris, C. M. 1953. A study of the building blocks of speech. JAcS. 25. 962-969.

Harris, K. S., G. F. Lysaught, and M. M. Schvey. 1965. Some aspects of the production of oral and nasal labial stops. L & S. 8.135-147.

Heffner, R-M. 1950. General phonetics. Madison: Univ. of Wisconsin Press.

Henke, W. 1967. Preliminaries to speech synthesis based upon an articulatory model. Conference preprints, 1967 Conference on Speech Communication and Processing, 170-177. Bedford, Mass.: Air Force Cambridge Research Laboratories.

Hiki, S. 1970. Control rule of the tongue movement for dynamic analog speech synthesis. JAcS. 47.85.

Hiki, S. and R. Harshman. 1969. Speech synthesis by rules with physiological parameters. JAcS. 46.111.

Hiki, S. and J. Oizumi. 1967. Controlling rules of prosodic features for continuous speech synthesis. Conference preprints, 1967 Conference on Speech Processing, 23-26. Bedford, Mass.: Air Force Cambridge Research Laboratories

Hiki, S., R. Ratcliffe, S. Hubler, and P. Metevelis. 1968. Notes on LASS circuitry. Working papers in phonetics 10, 12-41. Los Angeles: UCLA Phonetics Laboratory.

Holmes, J. N. 1961. Research on speech synthesis carried out during a visit to the Royal Institute of Technology, Stockholm, from November 1960 to March 1961. Report JU 11-4. Eastcote (England): Joint Speech Research Unit.

Holmes, J. N., I. G. Mattingly, and J. N. Shearme. 1964. Speech synthesis by rule. L & S. 7.127-143.

Honda, T., S. Inoue, and Y. Ogawa. 1968. A hybrid control system of a human vocal tract simulator. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, 175-178. Tokyo: International Council of Scientific Unions.

Ichikawa, A., Y. Nakano, and K. Nakata. 1967. Control rule of vocal-tract configuration. JAcS. 42.1163.

Ichikawa, A. and K. Nakata. 1968. Speech synthesis by rule. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, B171-4. Tokyo: International Council of Scientific Unions.

Iles, L. A. 1967. M. A. K. Halliday's "Tones of English" in synthetic speech. Work in progress 1, 24-26. Edinburgh: Department of Phonetics, Edinburgh University.

Iles, L. A. 1969. Speech synthesis by rule. Work in progress 3, 23-25. Edinburgh: Department of Phonetics and Linguistics, Edinburgh University.

Ingemann, F. 1957. Speech synthesis by rule. JAcS. 29.1255.

Ingemann, F. 1960. Eight-parameter speech synthesis. JAcS. 32.1501.

Kacprowski, J. and W. Mikiel. 1968. Recent experiments in parametric synthesis of Polish speech sounds. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, B191-4. Tokyo: International Council of Scientific Unions.

Kato, Y., K. Ochiai, and S. Azami. 1968. Speech synthesis by rule supplementarily using natural speech segments. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, B199-202. Tokyo: International Council of Scientific Unions.

Kelly, J. L. and L. J. Gerstman. 1961. An artificial talker driven from a phonetic input. JAcS. 33.835.

Kelly, J. L. and C. Lochbaum. 1962. Speech synthesis. Proceedings of the Speech Communications Seminar, paper F7. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology.

von Kempelen, W. R. 1791. Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine. Vienna: J. B. Degen.

Kenyon, J. S. 1950. American pronunciation. 10th ed. Ann Arbor: Wahr.

Kim, C-W. 1966. The linguistic specification of speech. Working papers in phonetics 5. Los Angeles: UCLA Phonetics Laboratory.

Koenig, W. H., H. K. Dunn, and L. Y. Lacey. 1946. The sound spectrograph. JAcS. 18.19-49.

Kozhevnikov, V. A. and L. A. Chistovich. 1965. Rech' artikuliatsia i vospriiatie. Translated (1966) as Speech: Articulation and perception. Washington: Joint Publications Research Service.

Ladefoged, P. 1964. Some possibilities in speech synthesis. L & S. 7.205-214.

Ladefoged, P. 1967. Linguistic phonetics. Working papers in phonetics 6. Los Angeles: UCLA Phonetics Laboratory.

Lawrence, W. 1953. The synthesis of speech from signals which have a low information rate. Communication theory, ed. by W. Jackson, 460-471. London: Butterworth.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. 1967. Perception of the speech code. Psychol. Rev. 74.431-461.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. J. Exper. Psychol. 53.358-368.

Liberman, A. M., K. S. Harris, J. A. Kinney, and H. Lane. 1961. The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. J. Exper. Psychol. 61.379-388.

Liberman, A. M., F. Ingemann, L. Lisker, P. C. Delattre, and F. S. Cooper. 1959. Minimal rules for synthesizing speech. JAcS. 31.1490-1499.

Lieberman, P. 1967. Intonation, perception and language. Cambridge, Mass.: M.I.T. Press.

Liljencrants, J. C. W. A. 1968. The OVE III speech synthesizer. IEEE Trans. Audio. 16.137-140.

Lindblom, B. 1963. Spectrographic study of vowel reduction. JAcS. 35. 1773-1781.

Lisker, L. and A. S. Abramson. 1967. Some effects of context on voice onset time in English stops. L & S. 10.1-28.

Lisker, L., F. S. Cooper, and A. M. Liberman. 1962. The uses of experiment in language description. Word 18.82-106.

MacNeilage, P. F. and J. L. DeClerk. 1969. On the motor control of coarticulation of CVC monosyllables. JAcS. 45.1217-1233.

Matsui, E. 1968. Computer-simulated vocal organs. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, B151-154. Tokyo: International Council of Scientific Unions.

Mattingly, I. G. 1966. Synthesis by rule of prosodic features. L & S. 9.1-13.

Mattingly, I. G. 1968a. Synthesis by rule of General American English. Supplement to status report on speech research. New York: Haskins Laboratories.

Mattingly, I. G. 1968b. Experimental methods for speech synthesis by rule. IEEE Trans. Audio. 16.198-202.

Mattingly, I. G. In press. Synthesis by rule as a tool for phonological research. L & S.

Mattingly, I. G. and A. M. Liberman. 1969. The speech code and the physiology of language. Information processing and the nervous system, ed. by K. N. Leibovic, 97-117. Berlin: Springer.

Mermelstein, P. In press. Computer simulation of articulatory activity in speech production. Proceedings of the International Joint Conference on Artificial Intelligence, Washington, D. C., 1969, ed. by D. E. Walker and L. M. Norton. New York: Gordon & Breach.

Moffitt, A. R. 1969. Speech perception by 20-24 week old infants. Paper presented to the Society for Research in Child Development, Santa Monica, Calif., March 1969.

Munson, W. A. and H. C. Montgomery. 1950. A speech analyzer and synthesizer. JAcS. 22.678.

Nakata, K. and T. Mitsuoka. 1965. Phonemic transformation and control aspects of synthesis of connected speech. J. Radio Res. Labs (Tokyo) 12. 171-186.

O'Connor, J. Desmond, and G. F. Arnold. 1961. Intonation of colloquial English. London: Longmans.

Öhman, S. E. G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. JAcS. 39.151-168.

Öhman, S. E. G. 1967. Numerical model of coarticulation. JAcS. 41. 310-320.

Oizumi, J., S. Hiki, and Y. Kanamori. 1967. Continuous speech synthesis from phonemic symbol input. Reports of the Research Institute of Electrical Communication 19.241-245. Sendai, Japan: Tohoku University.

Paget, R. 1930. Human speech. London: Routledge and Kegan Paul.

Peterson, G. E. and H. L. Barney. 1952. Control methods used in a study of the vowels. JAcS. 24.175-184.

Peterson, G. E., W. S-Y. Wang, and E. Sivertsen. 1958. Segmentation techniques in speech synthesis. JAcS. 30.739-742.

Pike, K. L. 1945. The intonation of American English. Ann Arbor: Univ. of Michigan Press.

Rabiner, L. R. 1967. Speech synthesis by rule: An acoustic domain approach. Bell System Tech. J. 47.17-37.

Rabiner, L. R. 1968. Digital-formant synthesizer for speech-synthesis studies. JAcS. 43.822-828.

Rabiner, L. R. 1969. A model for synthesizing speech by rule. IEEE Trans. Audio. 17.7-13.

Rao, P. V. S. and R. B. Thosar. 1967. SPEECH: A software tool for speech synthesis experiments. Technical Report 38. Bombay: Tata Institute for Fundamental Research.

Rosen, G. 1958. Dynamic analog speech synthesizer. JAcS. 30.201-209.

Saito, S. and S. Hashimoto. 1968. Speech synthesis system based on interphoneme transition unit. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, B195-198. Tokyo: International Council of Scientific Unions.

Scott, R. J., D. M. Glace, and I. G. Mattingly. 1966. A computer-controlled on-line speech synthesizer system. Digest of technical papers, 1966 IEEE International Communications Conference, 104-105. Philadelphia: IEEE.

Shankweiler, D. and M. Studdert-Kennedy. 1967. Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19.59-63.

Shearme, J. N. and J. N. Holmes. 1962. An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1-formant 2 plane. Proceedings of the Fourth International Congress of Phonetic Sciences, ed. by Antti Sovijatvi and Pentti Aalto, 234-240. The Hague: Mouton.

Stevens, K. N. and A. S. House. 1955. Development of a quantitative description of vowel articulation. JAcS. 27.484-493.

Stevens, K. N., S. Kasowski, and C. G. M. Fant. 1953. An electrical analog of the vocal tract. JAcS. 25.734-742.

Stowe, A. N. and D. B. Hampton. 1961. Speech synthesis with pre-recorded syllables and words. JAcS. 33.810-811.

Tatham, M. A. A. 1969a. Experimental phonetics and phonology. Occasional papers 5, 14-19. Colchester: University of Essex Language Centre.

Tatham, M. A. A. 1969b. On the relationship between experimental phonetics and phonology. Occasional papers 5, 20-26. Colchester: University of Essex Language Centre.

Tomlinson, R. S. 1965. SPASS--an improved terminal-analog speech synthesizer. JAcS. 38.940.

Troubetzkoy, N. S. 1939. Principes de phonologie. Tr. 1949 J. Cantineau. Paris: Klincksieck.

Umeda, N., E. Matsui, T. Suzuki, and H. Omura. 1968. Synthesis of fairy tales using an analog vocal tract. Reports of the 6th International Congress on Acoustics, ed. by Y. Kohasi, B159-162. Tokyo: International Council of Scientific Unions.

Vanderslice, R. 1968. Synthetic elocution. Working papers in phonetics 8. Los Angeles: UCLA Phonetics Laboratory.

Wang, W. S-Y. and C. J. Fillmore. 1961. Intrinsic cues and consonant perception. JSHR. 4.130-136.

Werner, E. and M. Haggard. 1969. Articulatory synthesis by rule. Speech synthesis and perception, Progress report 1, 1-35. Cambridge: Psychological Laboratory, University of Cambridge.

Wheatstone, C. 1837. Review of On the vowel sounds, by Robert Willis, Le mecanisme de parole, by Wolfgang v. Kempelen (= v. Kempelen, 1791), and Tentamen coronatum de voce, by C. G. Kratzenstein. London & Westminster Rev. 6 and 28.27-41.

Young, R. W. 1948. Review of U. S. Patent 2,432,123, Translation of visual symbols, R. K. Potter, assignor (9 December 1947). JAcS. 20.888-889.

149/150

On Time and Timing in Speech*

Leigh Lisker+
Haskins Laboratories, New Haven

However the student of language chooses to regard the object of investigation, whether as an assemblage of sentences[1] of indeterminate number or as a finite system of rules for sentence generation, he must at some point in specifying a language talk about elements and their arrangements. And however complex the network of relations that fix the elements within the sentence, the arrangement of these elements is the simple one of serial ordering. In the case of sentences "actualized" as pieces of speech, this serial ordering is necessarily one of temporal sequence, the elements of which are phonetic segments or "speech sounds," i.e., the lowest-level phonological units susceptible of physical description. This description is, in general, resolvable into a set of specifications with respect to a certain number of parameters, so that any two phonetic segments are comparable as points located within some multidimensional physical space. These parameters, whether they specify states of the speech-generating mechanisms, acoustic properties of the speech signal, or even the instructions to the speech apparatus, are usually chosen for their usefulness in accounting for listeners' ability to decide consistently that some speech pieces are repetitions of a single sentence and that others are instances of different sentences. Thus, for the linguist in particular, the problem of describing the phonetic segments of a spoken sentence is limited in that his interest is restricted to the potentially distinctive properties of those segments. His physical description of a segment is then partial and relational, constructed with an eye to the total ensemble of segments required by a universally applicable system of phonetic transcription. To be sure, the phonetician is in general also much concerned with the differential properties of segments, but his interest in speech behavior is not confined to those aspects directly relevant to the task of developing an efficient system for spelling or classifying sentences. For him it is not nearly enough to describe a sentence, here considered a class of speech pieces, as the mere encipherment of some particular graphical string, nor is he prepared to equate the physical description of a sentence with the sequence of physical specifications of the segments said to compose it.

To the student of speech behavior the object of interest in its most directly observable form presents itself as an ensemble of auditory signals generated by

---

*Chapter prepared for Current Trends in Linguistics, Vol. XII, Thomas A. Sebeok, Ed. (The Hague: Mouton).

+Also, University of Pennsylvania, Philadelphia.

[1] The term "sentence" might be understood to include sequences of sentences that constitute partial or complete texts.

complexly interrelated activities of body parts which control the movement of air into and out of the vocal tract. Whether considered in its acoustic or in its articulatory aspects, a speech piece is describable as a "time function" of one or more dimensions of the kind called "quasi-continuous." By this is meant that the signal, although not devoid of discontinuities or abrupt changes with time, resists efforts to analyze it into segments which are simply related to the phonetic segments established by the linguist's auditory-phonetic analysis (Lisker, 1957b; Fant, 1962; Lüdtke, 1969, 1970). For the linguist it has been said to occasion no difficulty that the speech signal does not yield physical segments matching those that underlie his phonological analysis (Chomsky and Halle, 1968: 294). His interest in the speech signal stops effectively at the level of auditory analysis,[2] for this provides a sufficient basis for developing a useful spelling system by which to represent those facts about speech that engage his concern. Moreover he may justifiably assume that his phonetic segments are ultimately relatable to the physical signal, whatever the difficulties that the phonetician might have to contend with in establishing those relations. In assuming this onerous, but not unrewarding, assignment, however, the phonetician need not take it to mean that his role is merely one of validating the linguist's phonetic description, on pain of being declared linguistically "irrelevant" in the event of failure. Beyond accepting the linguist's phonetic transcription as representing the result of some kind of perceptual-linguistic processing of the auditory signal, the phonetician is obliged neither to consider very seriously the physical content of the linguist's segmental description nor to adopt his alphabetic model as the most appropriate one for speech. Moreover the phonetician may well take issue with any implication that the investigation of speech behavior, except for the necessary minimum represented by the linguist's phonetic transcription, is devoid of serious linguistic interest (Mattingly and Liberman, 1970).[3] If the aim of phonology is to give a coherent account of the linguist's auditory analysis and its phonetic component is only just physical enough in reference to serve as an "objective" description of the elements of that auditory analysis, then the relation between the auditory unit and its physical specification is the simplest possible. It is also relatively uninteresting as a theory of how auditory signals recognized as speech are in fact identified with sentences of a language.

Underlying the linguist's graphical representation of a sentence is a model of the speech piece as a temporal sequence of articulatory states, their acoustic resultants or their neural-command antecedents, which are themselves largely "timeless" (Abercrombie, 1967:42, 80-81), in that a particular segment is no more to be characterized by the time interval over which its defining physical properties are maintained than its graphical representative is by the space it may occupy

---

[2] Whatever extra-phonetic information he brings to bear in deciding whether or not he has missed something in his phonetic description, that description itself is a matter of auditory decisions. The use made of grammatical information is a matter of "discovery procedure."

[3] The basis for this disinterest on the linguist's part may have been expressed most overtly by Hockett (1955:180), who supposed that no matter where in the communication channel we chose to study the linguistic signal, we should determine a set of elements that, despite very different physical properties, would show interrelations identical with those derived by the usual linguistic procedures. In short, the linguist would learn nothing he did not already "know."

on the line of print. Given the observed "elasticity" of speech in the time domain (Gaitenby, 1965), it is clear that the linguist's representation of a sentence must very particularly eschew any reference to absolute time intervals. What is invariably important about a segment in relation to time is not its temporal extent but its position in the temporal sequence of segments. The segments with their phonetic specifications might be thought of as samplings of the speech signal taken at a number of points in time, with the number for a given sentence being largely independent of the time taken for any particular performance of the sentence. On the other hand, however, it is recognized that speech events do, in point of physical fact, occupy some measurable time interval, for there are occasions when linguists include a reference to relative length in specifying a segment. Thus some vowels in a given language may be said to be shorter or longer in certain contexts than elsewhere, or they may show length differences in conjunction with differences in articulatory position. Such temporal differences are usually taken to be linguistically nondistinctive, an evaluation which the linguist may base on a particular phonological analysis but which the phonetician seeks to derive from their phonetically redundant status as the physical or physiological consequences of other properties of the speech contexts in which they have been observed (Abramson, 1962:109-110; Elert, 1964: 39-43; Hadding-Koch and Abramson, 1964). There is an extensive literature along these lines, reviewed most recently by Lehiste (1970:18-41), which reports various kinds of "conditioned" durational differences among classes of phonetic segment types in a number of languages. These differences are not only not significant linguistically, but many of those described in the phonetic literature are unreported in the linguist's phonetic description and are presumably not detectible by ear as a temporal phenomenon, if at all. That temporal differences may sometimes figure more importantly in distinguishing segment types is suggested, for example, by statements to the effect that flapped consonants and semivowels "admit of no duration" (Catford, 1968:330), presumably as distinct from stops and vowels, respectively, and that consonants that may function as syllable nuclei do so chiefly by virtue of a significantly greater duration (Jones 1950:139). Finally, linguists sometimes include relative length as a phonetic feature of segments in recognition of the fact that certain linguistic contrasts depend essentially on a speaker's maintaining a particular articulatory posture for a shorter or longer time interval (Abramson, 1962; Liiv, 1961, 1962; Nasr, 1960; Obrecht, 1965). In this last situation, where a temporal dimension would seem unavoidably to belong to the set of segmental properties, a common practice of linguists is to analyze a "long" segment into a sequence of identical "short" ones (Bloomfield, 1933:109-110; Hockett, 1955:77), whereby the need to include length as a segmental feature is obviated. This is not invariably done, to be sure, in which case length is then recognized either as a segmental property or as a special kind of segment, which might be defined phonetically as the prolongation of the articulatory position of an immediately preceding element. The problem of choosing among alternative "solutions" for representing a distinctive feature of length seems to be settled arbitrarily by the application of criteria that are not overly compelling (Gleason, 1961:282-284; Lehiste, 1970:44-46) and, in any case, do not rest on a secure phonetic basis (Jones, 1950:115-120; Hockett, 1955:76-79). Nevertheless, depending on the choice made, a feature of relative length is, or is not, included among the set of segment-specifying features.

Despite the options available to the linguist in deciding how to define the phonological status of contrastive length, there would seem to be reason to consider a feature of this kind a foreign intruder within the set of dimensions used to characterize the phonetic segments. Phonetic segments are, after all, thought of as the building-blocks of a static-phonetic and not a dynamic-phonetic description. If a certain utility is conceded the kind of description that restricts itself to the physical specification of the segment inventory and if, at the same time, one recognizes the need for a more dynamic phonetics not so closely tied to the segment, then it would seem that temporal features, insofar as they are interpreted perceptually as length, belong properly to the second type of description, whose domain is generally speech pieces of greater than unit segment composition. Unlike features of articulatory configuration, which, by and large, are considered characteristic of a segment in that they can be specified independently of context, temporal properties of a segment cannot, for the most part, be defined except as that segment participates in a longer piece of speech. The better way to put it, perhaps, is to say that length is not at all a property of segments as items in an inventory, but rather of longer entities which are in part describable as sequences of segments. In short, such temporal features as relate to length are prosodic in nature (Fry, 1968:370; Lehiste, 1970:2) in that their identification depends on the comparison of segments under conditions that exclude the possibility that the length differences have no sentence-differentiating function. Like features of differential pitch and prominence, relative length is also prosodic, according to Lehiste, in that it is "a secondary, overlaid function of inherent features" (Lehiste, 1970:2; also Peterson and Shoup, 1966:114). But perhaps the three features are not equally prosodic. Pitch and prominence differences function in a more nearly orthogonal relation to the sequence of segments than does length, and this may help to explain why it is the last which is often treated as just another linear segment. Then, in terms of a purely segmental description, the length of a speech piece is defined as the number of segments into which it is analyzed. If a linguistic contrast appears to depend on a length difference between two otherwise similar segments, the common linguistic practice of identifying the prolonged articulatory posture as a sequence of identical segments amounts to representing the length difference as a difference in the number of segments composing the contrasting speech pieces. Such an analysis would seem to recommend itself on the grounds earlier mentioned, quite apart from whether or not it would be "allowed" by the phonotactics of the specific language.[4]

Articulatory postures prolonged to a significant degree are not always treated as sequences of like segments, so that a feature of relative length may have the status of a segmental property. As such, it is, of course, not considered to represent a continuously variable quantity; rather does length come in quanta or "morae" (Bloomfield, 1933:110; Hockett, 1955:61). A short segment differs from a long one, generally, in that it contains one mora as against two or, in rare cases, three morae.[5] The notion of the mora, dictated by the convention

---

[4] Thus, for example, Lehiste (1970:44) allows this analysis in the case where a language has "consonant clusters that function in the same manner as long consonants...regardless of whether it is possible to demonstrate, phonetically, their geminate nature." One is entitled to wonder why it could not be said of a language devoid of consonant, or vowel, clusters that the only allowable consonant, or vowel, sequences are those which involve identical elements.

[5] Bloomfield in fact violates the usual practice of dealing in only integral numbers of morae (1933:111) and thereby most strongly implies the quantal nature of the unit. But the mora also appears to be not clearly different from a "degree of quantity" (Lehiste, 1970:48), so that its quantal nature is not always evident. Jones's "chrone" (1950:126) is, on the other hand, simply the actual measured duration and as such is continuously variable.

that speech be represented quantally, is consistent with, and complementary to, that of the "intrinsic duration" of segments (Lehiste, 1970:18-19, 27-30), which, for a given overall speech rate, is determined by their specific segmental properties. Where the length of a segment claims the linguist's special attention, this serves to alert him either to some contextually determined perturbation of its intrinsic duration or to the need for further segmentation of the sentence. Thus, an interval of the speech signal for which no change of articulatory configuration may serve to mark a segmentation point is nevertheless divided into two or three segments on a purely temporal basis--a judgment that the interval of articulatory stability occupies, not the one mora its segmental properties "entitle" it to, but two or three times that intrinsic duration. Whatever the phonetician may say about differences in segment size (Abercrombie, 1967:40), the segment as a unit of linguistic description is, then, most often specified without a temporal dimension, i.e., it is simply understood to be of unit length. The variability in segment length that may be observed in speech is then ascribed to a number of factors whose durational effects can be then separately determined (House, 1961:1174-1178; House and Fairbanks, 1953:105-113).

We see, then, that the linguist tends to minimize the temporal aspects of speech, treating it as an orderly sequence of evenly spaced units of fixed size. It may be said that this picture corresponds to the linguist's purpose in formulating his phonetic specification, which is to classify sentences. For this it suffices to name the segments as sequentially ordered components of sentence designations, using as names for the segments sets of distinctive features by which the segments may be located as points in orderly array within some physical, phonetic, or phonological space. From the standpoint of the phonetician, the linguist's purpose is an overly modest one, and the static representation that satisfies it is not to be considered an adequate description either of speech activity or of the acoustic signals it generates. The segment properties, which the linguist asserts have distinctive functions, must for the phonetician be placed in the context of a more general understanding of the rules governing the operations of the speech-producing and speech-perceiving mechanisms. Not least among these may well be timing rules which determine the movements of the various articulators and their temporal interrelations, for there is good reason to think that, in order to yield the proper order and number of segments required to produce a given sentence in a normal manner, perhaps more than one hundred muscles must be coordinated to operate within tolerances of sometimes less than a tenth of a second (Lenneberg, 1967:91-103). Failure to maintain relatively fixed phase relations among the component gestures of a given segment sequence will result in variation in the number of acoustic segments generated and sometimes in the number and character of the phonetic segments perceived (Kantner and West, 1960:264-269). There is also clear evidence that many kinds of deficient speech, including both humanly produced (John and Howarth, 1965; Lowe and Campbell, 1965; MacNeilage et al., 1967; Shankweiler et al., 1968) and machine-generated varieties (Mattingly, 1966), involve disorders of timing.

In the foregoing we have been considering the phonetic segment as an articulatory and/or auditory event that is relatively stable over some time interval. The temporal extension of the segment is also recognized in cases where its component features, theoretically co-occurrent, have certain manifestations that are noticeably not simultaneous. Here the temporal factor is reflected more in judgments of temporal order than extent. Thus, a segment will be described as a glottalized stop, with closures of both the glottis and the oral cavity, but the two releases may not be synchronous, so that their temporal order must also be specified (Hockett, 1955:35). It can be objected that such entities hardly qualify as phonetic segments, but the basis for this objection is by no means obvious, inasmuch as a reading of the literature reveals rather different notions of what the "phonetic segment" ought to mean. Much opinion on the matter is flatly assertive and to a considerable extent controversial because, as in other areas within the linguistic sciences, it fails fully to distinguish between matters of definition and matters of physical and perceptual fact. Perhaps the most purely phonetic definition of the segment is that of Pike (1943: 107), who described it as "a sound (or lack of sound) having indefinite borders but with a center that is produced by a crest or trough of stricture during the even motion or pressure of an initiator."[6] But in the literature generally, a distinction between the phonetic segment as an articulatory state and as a unit whose size is determined essentially by phonological considerations is not clearly made. There has been no lack of definitions, ranging from the purely physical one of Pike's, over the purely phonetic or physico-perceptual (Abercrombie, 1967:42), to definitions that make the phonetic segment purely phonological in nature (Gleason, 1961:248-249). It is undoubtedly true that in practice the linguist's phonetic segments are not to be equated with "prelinquistic," that is, purely auditory, judgments of articulatory states. Certainly Chomsky and Halle (1968:293-295) are correct in describing the linguist's phonetic segment, not as a judgment of physical fact uninformed by linguistic knowledge or intuitions, but rather as fundamentally a phonological unit described by a set of physical features that are themselves selected on grounds as much phonological as physical. For them, it is then a phonological unit, though it is at the same time unitary by some universally valid phonetic theory (1968:28). Perhaps because the phonetic segment, like the syllable, is one which linguists and other users of language seem to have strong convictions about, the definitions advanced have often taken the form of assertions as to what it is, rather than being proposals as to the kind of definition that will yield segments appropriate to some stated goal. Of course, to complicate matters, it is not true that either linguists or other people are always quite certain about what the segments of speech piece are, even when no definitional dispute is overtly involved. A notorious example is furnished by the history of dispute concerning the status of the affricates.

If the phonetic segment is characterized as a complex of simultaneous features, then there would appear to be no place in a phonetic representation for entities such as "prenasalized stops" (Jones, 1950:78-81; Ladefoged, 1964: 23-24) or "occluded nasals" (Bauernschmidt, 1965:477, 480-481), which can only be considered phonetically unitary if we can discover some basis for asserting that they constitute signal complexes which are integrated to a degree that ordinary sequences of nasal and stop consonants are not. Unlike entities such as the glottalized or aspirated stops, for which we may plausibly assume certain articulatory maneuvers to be executed in close, if not precise, synchrony,

---

[6] The "phone type" of Peterson and Shoup (1966:112) is perhaps even more stringently phonetic in nature, but so much so that it may not be synonymous with the phonetic segment of standard linguistic description. Thus, for example, while their system of definitions provides a place for "complex" phones, these do not apparently include a class of affricates, which certainly belong to the set of phonetic segments available to the linguist.

a prenasalized stop or an occluded nasal involves a sequence of two necessarily nonoverlapping states of velar opening. If the segment is to be associated with a single articulatory state, it would seem that such entities ought to be considered two-segment sequences. In the case of glottalized and aspirated stops, where the nature of the temporal relations among the component articulations is neither one of simultaneity nor one of strict succession, there seems to be no firm phonetic basis for deciding whether to regard them as single segments or as sequences. If the treatments of such phonetic complexes that are found in Hockett (1955) may be considered representative lignuistic practice, then clearly the notion of the phonetic segment as a purely phonetic entity is derivative; its limits are not so fixed that it can be said to represent a particular state of the vocal tract, but it is rather a phonologically unitary stretch of speech signal to which a phonetic description is attached. This phonetic description is most often, in fact, the specification of a single articulatory state, in which case the linguist's phonetic segment coincides with one to which the phonetician would as happily subscribe. But this need not be the case, and then it becomes necessary to describe certain segments as ordered sequences of articulatory states. That a speech piece will, by and large, be analyzed auditorily into segments that are very much the same, no matter what the language, is certainly true (Pike, 1943:116), and these will be of the size of phonological units. There nevertheless remain certain phonetic events whose status as segments is ambiguous except by some phonological decision criterion.[7] Where this leads to the establishment of complex phonetic segments of the kind just mentioned, a feature of temporal order must be included in the overall inventory of features needed for the phonetic description of segments generally. Then one might define as phonetic segments such entities as preaspirated stops (Heffner, 1950:168), along with the more usual postaspirated types, and suppose that these two classes are adequately distinguished by the different ordering of their constituent events. It would then seem, however, that the nature of this temporal feature is not different from the ordering relation among segments in longer sequences and that therefore any notion of the phonetic segment as a purely phonetic unit is strictly untenable.

Let us now turn to the question of the affricates. Laziczius (1961:62-66) gives a fairly detailed resume of the various opinions expressed concerning these items, and one gathers an impression of substantial agreement on their physical nature and endless wrangling as to their segmental status. They are produced, it seems fairly well agreed, quite in the manner of stop consonants insofar as they involve complete oral occlusion, but the time course of their releases is such that a noise interval results that is longer and is identifiable as a fricative homorganic with the stop element. The affricates are, then, simply slowly released stops (Heffner, 1950:120; Hockett, 1955:81; Gleason, 1961:248; Abercrombie, 1967:147). Their stop and fricative phases must necessarily be homorganic, and presumably the stop component can have no stop-like release intervening between the occlusion and the fricative phases, since the latter is, in fact, the release. This slower release in the affricated as against the ordinary stop consonant produces an interval of noise that for some scholars is audible as a brief fricative (Heffner, 1950:249; Abercrombie, 1967: 147), and for at least one other is "too short to be heard separately" (Gleason, 1961:249), at least, we may infer, by speakers of a language in which the affricate represents a single phonological unit. According to some accounts, moreover, the fricative phase of the affricate is distinguished from a true fricative again by a difference in duration (Gerstman, 1956, 1957; Brooks, 1964). Thus it appears that there is general agreement that the gesture of opening after a total occlusion is under temporal control to the extent that three rates of opening are distinguishable. As between the affricates and the stop + fricative sequences, it has been supposed that the first are released with a single articulatory movement, while in the sequences "the articulator goes through successive motions" (Hockett, 1958:81). However there seems to be solid evidence only for the durational differences, and none either to confirm or deny that these differences apply to articulatory movements that are otherwise identical.

---

[7] These are Pike's "fluctuants," which do not constitute a phonetic class, apparently, but are simply those events about whose segmental status linguists are in doubt.

The tenability of any purely phonetic basis for deciding on the number of segments composing an affricate is open to question no matter whether it is claimed to be temporal or articulatory or both together, for it is by no means certain that these aspects of speech activity can be quantified to yield just the classes to be defined. Thus, some phonetic descriptions suggest that releases of oral occlusion may show as many as four durations, by which stops, affricated stops, affricates, and stop + fricative sequences may be distinguished (Gimson, 1962:153-154,166-169; Malmberg, 1963:43,50-51). Moreover, the usefulness of the articulatory argument based on the necessary homorganicity of the affricate components is reduced when one finds it applied as a basis for justifying the inclusion of both [pɸ] and German [pf], and English [tɹ], as well as [tš], among this class of items (Gimson, 1962:31,168) and when it is recalled that the effects of coarticulation, under which this "homorganicity" may or may not be subsumed, have only just begun to receive quantitative treatment, in respect to both their magnitude and their temporal extent (Chistovich et al., 1965:126-132; Öhman, 1966:151-168; Öhman, 1967:310-320; Lindblom and Studdert-Kennedy, 1967:830-843). If a phonetic description of speech activity must be the description of segments in linear order that the linguist deals in, then precise models of coarticulation are needed before the segment can be taken as a phonetic, rather than a phonological, unit.[8] Until the notion of a phonetic segment as an interval over which a fixed articulatory position is maintained can be better reconciled with the observation that at any instant one or more of the speech organs is in motion (Gimson, 1962:42), it would be safer to pretend, following Gleason (1961:248), that "phonetically there is no basis for the kind of segmentation which we customarily use" than to insist that the elements that a segmentation yields can all be specified by a set of fixed articulatory configurations. Neither point of view can be completely valid, however; there are certain kinds of acoustic discontinuities that perhaps are always interpreted as segment boundaries, but not every signal change has this perceptual effect (Pike, 1943:46-47; Fant, 1960:21-26).

Despite the fact that they may define certain auditory complexes as phonetic segments, some linguists and phoneticians nevertheless seem averse to abandoning utterly the idea that their segments are fundamentally units derived by a segmentation that is entirely divorced from phonological considerations. For them the fact that their phonetic segments converge generally with those derived

---

[8] Purely phonological considerations most obviously motivate proposals such as those of Hofmann (1967), who would classify all initial clusters in English as unitary segments.

by a language-specific segmentation based on phonological analysis is not only independently significant, but it impels them to suppose that all phonetic segments must be phonetically unitary, despite any auditory complexity some of them might display. Thus, if considerations of coding efficiency indicate that certain stop + fricative sequences should be treated as units, they feel compelled to discover that phonetically their constituents are more intimately bonded than are stop + fricative sequences generally (Hockett, 1955:164). This would not eliminate the need to include a feature of temporal order in the description of complex segments but would allow us to distinguish phonetically, and most probably on a temporal basis, between these and sequences of stop + fricative segments. Thus, for example, a phonologically unitary sequence in Polish is distinguished from a sequence of segments presumably comprising the same phonetic events either on the basis of a difference between a close and an open transition from one event to the other (Bloomfield, 1933:119) or by a difference in the durations of the fricative intervals (Brooks, 1964). Similarly, a one-segment occluded nasal [n^t] differs from a sequence [nt] in that the nasal component of the first is longer, while the stop is shorter, less forcefully articulated, and more likely to be "contaminated" by voicing than are the corresponding events of the segment sequence (Bauernschmidt, 1965:480).

There is another strategy available for dealing with a sequence of phonetic events, which the linguist would regard as phonologically unitary and therefore to be distinguished phonetically from a sequence of segments. Instead of referring the sequence of events to a sequence of articulatory states which may be especially closely integrated when viewed as parts of a single segment, he can suppose that they are generated by the interplay of several articulatory maneuvers which are essentially synchronous (thus, Chomsky and Halle, 1968:326-329). Though this mode of phonetic accounting may not always remove a temporal dimension from the set of segment-specifying features and though it seems designed *ad hoc* to preserve the integrity of the phonetic segment as the linguist's basic unit of transcription, still it is of interest in that it allows for the possibility that the temporal order of acousticoperceptual events is not precisely matched by that of corresponding events at some deeper level of the speech-generating process (Lenneberg, 1967:93-98). It is intriguing to postulate that the perceptual unity of the phonetic segment, where it is acoustically a complex of successive events, reflects a unity at some level of the speech communication process that is less accessible to direct observation. In one sense linguists have presumably always been ready to believe this, i.e., that somewhere in the speaker's nervous system there are "units of internal flow"(Hockett, 1955:5) corresponding to the linguist's discrete phonological elements and perhaps to the native speaker's intuitions as well (Jones, 1950: 79). This belief, however, amounts to little more than an assertion of the "psychological reality" of such entities and not necessarily to an assumption that the combination of certain specifiable articulatory properties or movements, or perhaps neural commands, generates the observed complex of acoustic events or that the transformation of the segment as a complex of co-occurrent features at the deeper level into the not nearly so well integrated acoustic complex is in conformity with any generally statable rules (but see Pike, 1943: 114; Chomsky and Halle, 1968:324). Assertions as to features at some deeper level are by definition less easily tested by physical observation, but with the development of new laboratory techniques in recent years, hitherto inaccessible stages of the speech process are becoming vulnerable to study (see, for example, the papers by Harris and Sawashima in the present volume).

The role of a feature of temporal ordering is most obviously important in describing acoustically complex segments, but it cannot be assumed to be negligible in the case of the simple ones. This would amount to supposing that the nature of the articulatory movements and the acoustic signal produced during the first half of the interval "occupied" by a phonetic segment is essentially symmetrical with that in the second. The only immediately obvious motivation for such an assumption would, I think, derive from the persistent tendency to view the graphical representations of speech as an adequate model of the speech process. Since the alphabetic signs of a phonetic transcription are freely transposed and "spaceless," in the sense that no significance attaches to any asymmetry in their shapes, it might be supposed that the phonetic segments are themselves as freely movable. There is abundant evidence that this is just not so. Thus, for example, attempts to fabricate acceptable speech into "segment-sized" bits and reassembling them in different orders have regularly failed expectations based on the assumption that all tape snippets bearing the same phonetic segments were identical in their combining behavior (Harris, 1953; Liberman et al., 1967:441). Nor can a speech piece consisting of auditorily simple segments be recorded and, upon being replayed with reversal in time, be heard as the same segments in a sequential order the reverse of the original (Harrell, 1958). Recent experiments in the perception of dichotically presented speech stimuli also indicate that when these are perceived as single, "fused" speech pieces, the perceived order of segments is not generally predictable from the temporal order of presentation, but is significantly an effect of their nonuniformity in the time domain (Day, 1970). Thus, segments of the acoustic speech signal, selected as physical correlates of phonetic segments, are not freely commutable elements, for their shapes make them, in general, compatible only with their immediate neighbors in the speech piece in which they originated, plus perhaps certain others not predictable on the basis of the phonetic transcription (Harris, 1953:962). In other words, the phonetic segments are acoustically shaped not like similar beads on a string, but rather like the pieces of a jigsaw puzzle.

So far in the present discussion, we have been playing the phonetics game as the linguist plays it, taking as the primary data his auditory phonetic evaluations of speech and seeking out physical properties that might correlate significantly with them. Because auditory analysis, both with and without the light of linguistic knowledge, gives us speech in the form of phonetic segments in temporal order, we look for discontinuities in the acoustic signal that will mark off time intervals matching the phonetic segments in some sense. If a piece consists, let us say, of x number of phonetic segments, then we seek x + 1 points (including those of onset and termination of the piece) at which to cut the physical signal. The resulting x time intervals are then paired with the same number of phonetic segments in such a way that we feel represents an optimal matching of the two kinds of elements. The method involves a knowledge of the acoustic correlates of certain articulatory events that can serve as temporal reference points in the acoustic signal, e.g., the brief noise "burst" of a stop release, together with an ability to follow variations in the acoustic signal that reflect the cyclic changes in oral aperture at syllabic rate, as well as the gross differences which correspond to alternations in voicing state. But this matching procedure, however "optimal" it be judged, does not guarantee the significance of the pairing relation it establishes. The assumption that somewhere in the acoustic signal are to be found properties which can be shown to correspond to those ascribed to the phonetic segments by no means implies that the acoustic signal is necessarily to be segmented along the time dimension as described above, for this procedure presupposes that the signal must be parceled out among the segments with neither overlap nor residue. Every bit of acoustic signal must, in short, be assigned to one and only one segment. Since there is an abundance of evidence to refute any such simple view of the relation between the phonetic evaluation of speech and its acoustic properties, the meaning of the kind of segment matching we have just described would seem at best to be not transparently clear. There is no problem about segmenting the acoustic signal, since, contrary to the once fashionable emphasis on the continuous nature of speech, we usually find more acoustically reasonable cutting points than we know what to do with (Fant, 1962:30). At the auditory level, however, it is more questionable that we may speak of boundaries "between" phonetic segments (Pike, 1943:107). If one nevertheless accepts a view of speech as a flow of discrete segments that follow one another without hiatus, assumes this view to be applicable without qualification to the acoustic signal, and performs the matching operation, there still remains the far from trivial matter of deciding what aspects of each segment are to be isolated as physical correlates of the matched phonetic segment. This problem is of course central to the field of acoustic phonetics, and knowledge of the phonetically significant features is by now fairly detailed (Delattre, 1968). The determination of these features has been a fairly straightforward matter, particularly since the development of techniques for the controlled synthesis of speech-like auditory signals, for the set of elements to be specified physically is well defined by auditory judgment and linguistic function. The physical specifications insofar as possible are correlates of the static-phonetic descriptions of the phonetic segments, i.e., vowels and certain of the consonants are described by fixed "target" values of particular acoustic dimensions, and the values exhibited by these segments in running speech are viewed as more or less systematic deviations from those targets (Stevens et al., 1966; Lindblom and Studdert-Kennedy, 1967). The precise location of acoustic segment boundaries is not crucial to this enterprise, at least with regard to the specification of the simple phonetic segments, although the classification of a segment as "simple" or "complex" would naturally depend on the placement of those boundaries.

Matters are rather different when we come to consider the question of matching acoustic and phonetic segments with respect to the time dimension. In the first place, as we have seen, the feature of length is only marginally considered to be a property of the phonetic segment. If an articulatory position seems to be maintained for different durations in linguistically distinct speech pieces, the linguist does not promptly consider length a segmental feature; he is more likely to look for some other phonetic difference by which to explain the time difference. Whereas segments are classified regularly with respect to such features as voicing, nasality, tongue height, and lip posture, there is no purely phonetic classification of segments on the basis of their perceived durations. Only where such a classification is forced by the recognition of length contrasts in a particular language is this done, and then only after the linguist has searched in vain for other feature differences that might account for the temporal difference. This reluctance on the part of linguists to recognize a feature of temporal control at the segmental level is most strikingly exemplified in Chomsky and Hall (1968:317,318,327). Thus, in general, phonetic segments can hardly be said to have a well-defined status with respect to the dimension of time. Of course, since we here speak of phonetic segments, we are thereby considering time in the sense of a perceptual dimension, one that is independent of phonological considerations. In the second place, the determination of segment boundary locations in the acoustic signal is of obvious importance if one wants to make physical time measurements that can be meaningfully related to the length, or perceived duration, of segments or longer stretches. Moreover, unlike the situation with respect to the indisputably segmental features, it is not generally possible to specify target values for the duration of a phonetic segment, since its so-called intrinsic duration can be only one factor, and most probably not a major one, in determining its duration in any given context, as produced by a given speaker, in a particular text, and under particular social and physiological circumstances. It is, of course, possible to define classes of acoustic segments as physical correlates of the phonetic segments and to proceed to determine their mean durations, but it seems farfetched in the extreme to suppose that these values are targets in the same way that the formant frequencies of isolated vowel sounds are construed to be (Lindblom, 1967:4-5).

It is perhaps useful here to emphasize a distinction generally made, and which we have been observing more or less consistently, between duration as a physically measurable attribute and length as its closest single analogue in the domain of perception (Malmberg, 1963:122). This distinction, which parallels others drawn between physical dimensions and their perceptual correlates (e.g., fundamental frequency vs. pitch) is perhaps less carefully observed than those others because the relation between duration and length is felt to be so much closer. Fundamental frequency, the number of fundamental periods per second, is a measure of glottal pulse rate, but pitch is not simply a subjective estimate of that rate. Pitch is rather an auditory property of a pulsed acoustic signal whose frequency is such that the individual pulses are not separately perceived. Length, on the other hand, is an auditory judgment as to the duration of a sound signal, where this duration presumably might be more accurately measured by instrument. Pitch is, by its nature, not subject to instrumental measurement. Of course, even with respect to duration and length, particularly when dealing with speech signals, we can only say that the accuracy of statements about length may be determined simply by instrumental measurements of duration if we accept the listener's description of his length judgment as an estimate of temporal duration. Now while we should in fact expect the duration of an auditory signal to be the major determinant of its length, the possibility exists that the relation is nonlinear (Fry, 1968:386), and that factors other than simple duration are of significance. Moreover there still remains the "problem of segmentation," which, in the present context, is one of deciding what acoustic segment or sequence it is whose duration should be defined as the physical correlate of the length of a given phonetic segment. This requires some decision if phonetic segments are to be said to have temporal extension in the sense of physical duration. Failure to define explicitly the physical intervals makes for difficulty in understanding exactly what is meant by published statements giving precise (within tolerances of a few milliseconds, apparently) durations of various segment types. One cannot simply carry out measurements of the durations of selected intervals in spectrograms and report these as durational values for vowels, stop consonants, or other segment types; we do not see vowels or stop consonants in the spectrogram, we only find evidence of them. The durations reported are of acoustically defined sections of the acoustic signal, and not strictly of phonetic segments, even as measurements of fundamental frequency are not determinations of pitch. These considerations, obvious enough, have prevented neither instrument manufacturers from marketing "pitchmeters" nor phoneticians from devising conventions that allow them to refer to the durations of phonetic segments.

In adopting measuring conventions for defining, and then measuring, the duration of vowels, which is the class of segments that has enjoyed most of this kind of attention, it has not, as far as I know, been clearly stated what the basis for selecting the acoustic markers of vowel onset and termination should be. The onset of vowels has been fixed sometimes at the point where the glottally excited, full formant pattern is established following silence or the release of a preceding consonant constriction (House and Fairbanks, 1953:107; Abramson, 1962:29), sometimes at the point of the consonant release (Peterson and Lehiste, 1960:694) or at a point where the "intensity curve rises sharply" (Fischer-Jørgensen, 1964:182) or where the formant frequencies judged appropriate to the vowel are attained following the transitional movement from a preceding consonant (Delattre, 1964). Hadding-Koch and Abramson (1964:98) included transitions within the acoustic interval defined as the vowel. Fónagy and Fónagy (1966:15) adopted the arbitrary procedure, but really no more arbitrary than any other, of grouping doubtful intervals with the immediately preceding acoustic segment. The vowel termination in most cases was located at a point corresponding to the end of the transitional movement to the constriction of a following consonant, though Delattre (1964), consistent with his view of transitions as essentially consonant cues, determined vowels at a point where the formants enter the transition phase.

Where the object of acoustical measurements is to specify "consonant duration" the conventions are in general complementary to those for the vowels. The literature on measurements associated with consonant length is not large, and perhaps almost the whole of it identifies as the appropriate objects of measurement the acoustic segments that correspond to intervals of constriction. In the case of certain of the languages for which durational data have been reported, i.e., English, German, Swedish, and Danish, all languages which include voiceless aspirated stops in their segment inventories, the decision as to the status of the aspiration phase might well affect the magnitudes of reported consonant durations by as much as 75 percent (Lisker and Abramson, 1964).

In view of the somewhat different measuring conventions followed by various researchers, to say nothing of the fact that some reports fail to describe any in satisfactory detail (Parmenter and Treviño, 1935:129-133; Fintoft, 1961:19-39), the comparability of their data is seriously to be questioned (Laziczius, 1961:119), certainly so far as the absolute magnitudes reported. Aside from the matter of measuring conventions, it is known that individual speakers show large differences in overall speaking rate and that at higher rates it is not the case that segments are eliminated to the point where those retained preserve the durations exhibited at slower rates (Gaitenby, 1965:3-4).

The variability in the criteria applied in segmenting the acoustic signal, which derives from the fact that the number and location of "natural" segmentation points in the acoustic signal do not conform in detail to an intuitively satisfying segmentation based on auditory judgments ('t Hart and Cohen, 1964:35), suggests the need for some criterion, not necessarily acoustic, by which to agree on a segmentation procedure. One possibility might be to select intervals for measurement that best match listeners' judgments of relative length. Thus, for example, on the question of whether the acoustic segment identified with aspiration should be considered a phase of the vowel or of a preceding stop, the decision would depend on which assignment yielded physical time values in better agreement with judgments of the relative length of the vowels following aspirated as against unaspirated stops. It might, alas, turn out to be a case of the halt leading the blind, if listeners simply proved unable to make consistent durational judgments when given this task. From figures given by Peterson and Lehiste (1960:701) for American English, it would seem that the durations at issue do not differ by much more than the general just-noticeable differences established for speech sounds (Lehiste, 1970:13). Moreover, when the durations of vowels following aspirated stops are determined both with and without aspiration as part of the vowel, the average of these two is almost exactly equal to the mean duration of vowels after /b,d,g/. In other words, given the likely event that listeners do not regularly report differences of vowel length that can be correlated with the aspirate/inaspirate difference in the preceding stop, we cannot simply settle the matter by showing that one assignment of aspiration will yield more nearly equal durations for the two sets of vowels than will the other.

One might return to the static-phonetic mode of description and elect as appropriate intervals those portions of the acoustic signal that, by some more or less arbitrary definition, can be considered to be "essentially steadystate" (Lehiste and Peterson, 1961:272; Gay, 1968:1571). The measured durations of such acoustic segments would then be defined as the durations of particular phonetic segments. If this procedure were consistently followed, the picture presented in the literature as to the relative durations of different segment types would require serious revision. Thus, statements to the effect that the vowels in English take up a disproportionately large part of total speech time (Parmenter and Treviño, 1935:130-131) would certainly turn out not true, given the likelihood that many vowels, the unstressed without doubt, are by any reasonable definition steadystate for, at most, half the durations often ascribed to them (Bernard, 1970:92; Lehiste and Peterson, 1961:275; Lindblom and Studdert-Kennedy, 1967:831). The well-established practice of equating the consonant with a relatively steadystate signal interval, e.g., the closure silence of a stop, while taking as the vowel an entire interval of voiced formant pattern, would seem to reflect the application of some highly arbitrary segmentation criterion, or at any rate one that, justifiable or not, has not been put into an explicit form. By the usual assignment of acoustic segments to phonetic, we have vowels that may be described as relatively long but rather unstable and consonants that tend to be shorter but with relatively fixed acoustic characteristics. If one takes the extensive data on American English supplied by Peterson and Lehiste (1960:702) and by Lehiste and Peterson (1961:275) to derive durations of steadystate stretches within the total intervals assigned the vowels, one arrives at values much more in line with those reported for consonant durations.

At this point, it is appropriate to ask whether it is necessary to follow the linguist's precisely segmental model of speech in the description of its temporal aspects. As Laziczius (1961:117) put it, "die Laute einer Lautreihe reihen sich nicht wie die Perlen eines Halsbandes nebeneinander, sondern greifen ineinander ein wie die Glieder einer Kette," and this has been amply demonstrated by the acoustic analysis of speech signals (Fant, 1962:30) and by experiments in their synthesis (Liberman, 1970a:309-315). We may believe, in accord with the segmental model, that the speaker really "wants" to produce a string of neatly discrete sound, or articulatory, units, and that only certain linguistically irrelevant constraints on his ability to execute these at the rates actually achieved in the production and transmission of phonetic segments interfere (Heffner, 1950:204-207; Fónagy and Magdics, 1960; Osser and Peng, 1964; Lenneberg, 1967:90; Liberman, 1970a:306). There is reason to believe, however, that the speech decoding mechanism which is presumed to operate on the auditory level could not handle discrete units at those rates (Liberman et al., 1967), and this can be true at the same time that it is also true that speech is normally slower than it need be for comprehension (Fairbanks et al., 1957; Fairbanks and Kodman, 1957; Klumpp and Webster, 1961; Endres, 1968; Sticht and

Gray, 1969). Even if the rate of segment production hovers about the lower frequency threshold for pitch perception, say between 10 and 20 segments per second (derived from Heffner, 1950:204-206), the fact that time compression of as much as 50 percent may not seriously reduce intelligibility ought to bring the rate of segment identification above that threshold, perhaps to as much as 30 per second (Liberman et al., 1967:432). If, at the same time, we bear in mind that listeners appear to be unable to judge accurately the temporal order of discrete auditory stimuli until their individual durations are as much as 700 msec (Warren et al., 1969:587), in the case of nonspeech sounds, and that sequences of vowel sounds are not correctly ordered until their durations are more than 100 msec each, it seems impossible to reconcile all this with the strictly segmental model of speech. Moreover, it becomes difficult to know what inferences to draw from the data published on the durations of the various classes of phonetic segments, or to suppose that these durations have any very close connection with the real units of speech production and/or perception.

If we follow Fant in the exercise of analyzing the spectrogram of a short piece of English (the piece is the short form "Santa Claus," Fant, 1962:30), we isolate nine components, one for each phonetic segment alleged to constitute the piece, which partially overlap in such a way that they yield twice the number of segments at the acoustic level. Most of these acoustic segments comprise acoustic features associated with two phonetic segments; of the remainder most provide cues to either three or four segments simultaneously. If we further ask how much time these nine acoustic components corresponding to the phonetic segments would occupy if they were purely sequential, so that instead of eighteen acoustic segments there were only nine, the answer is that the total duration of the piece would be increased by more than a third. If we were to make the extreme assumption that each of the overlapping components has the minimum duration required for auditory detection and proper identification of the corresponding phonetic segment, assuming that, in fact, they were clearly identifiable in the temporal order in which they were actually produced, then the durations of these components would seem the appropriate items to measure, since they represent more nearly the durations over which the acoustic reflexes of the separate gestures characteristic of the piece were present in the signal. There are certain further observations that may be made about this sample analysis, which in effect treats the phonetic segments as features of the acoustic signal. While the briefest acoustic segment is of the order of 10 msec in duration and the mean segment duration is about 115 msec, the shortest component is almost 60 msec long, and the mean component duration is greater than 150 msec. The sequential ordering of the phonetic segments is reflected in the fact that each component either begins or terminates before its successor, and never begins or terminates after its successor. Each component co-occurs, in its initial part, with one or more of its predecessors, and, later on, with one or more successors. Thus the time interval for a given component is acoustically far from steadystate and may even contain sharp discontinuities. We do not know how much of this complexity is actually required in order that the signal pass as intelligible speech, but some at least is indispensible. This means that the listener is capable of tending to more than one component simultaneously presented in the course of decoding the linguistic message (Liberman, 1970b:242-244). If Fant's sample analysis of "Santa Claus" correctly represents the way in which the acoustic cues are temporally distributed in speech generally, then one must say that only exceptionally does the listener encounter an acoustic segment which includes only a single component. In Fant's example, the initial and final acoustic segments are of this type, and otherwise only the phonetic segments [k] and [ɔ] are represented by components which are not always accompanied by components associated with other phonetic segments. The durations of the two one-component acoustic segments are about 55 and 200 msec for [k] and [ɔ] respectively, values that are rather smaller than some phonetic segment durations reported in the literature (e.g., Kent and Moll, 1969:1550; House, 1961:1174) but quite as large as others (e.g., Sharf, 1962:28). On the other hand, the total durations occupied by the components representing the [k] and [ɔ] segments are more like 200 and 340 msec respectively, values that are quite a bit larger than any reported for such segments. However, as we have noted before, a variety of factors affect the durational properties of speech, some drastically, so that it is difficult to evaluate the credibility of any magnitude report without certain information as to the nature of the speech sample measured and the conditions under which it was elicited.

If we elect to equate each phonetic segment of a speech piece with an acoustic segment in one-to-one fashion, choosing steadystate segments, and then define the duration of the acoustic segment as the physical correlate of phonetic segment length, we thereby say, in effect, that transient or transitional segments make no contribution to perceived duration. If, on the other hand, we make the component, i.e., the stretch of signal over which the acoustic cues to a phonetic segment are spread, the basis for determinations of duration that are to be related to length, we are led to the apparent absurdity of supposing that a single acoustic segment might contribute at once to the length of as many as four phonetic segments (as in Fant's example, cited above). Either measuring convention will lead to difficulty at some point. If measurement is restricted to steadystate segments, we must suppose that, perceptually, the duration of a speech stretch is determined by the summed durations of these steadystate segments, and this clearly is untenable. If, on the other hand, we choose the component, then we should reckon with another possibility, namely, that variation in degree of encoding or coarticulation with consequent change in total performance time might not be perceived as a variation in length at all, provided there were no change in the durations of the individual components. As a possibility, this latter is not so much implausible as a matter of unverifiable conjecture. But in the case where only steadystate segments are to be measured, it is difficult to believe that transitions make no contribution to length. It seems far more likely that they figure importantly in the length of vowels, for the degree of steadiness of an acoustic stretch associated with a vowel depends in part on the particular consonantal context, even where no length difference is perceived (see, for example, the spectrographic patterns in Öhman, 1966:160-162 and in Liberman, 1970a:309). Only if one believes that because transitional segments are generally more important for consonant that for vowel identification (Delattre, 1964, 1968) they must therefore be entirely assigned to the consonant, can one suppose that they contribute nothing to the perceived duration of the vowel. So far as duration is concerned, the contrary is more likely true: so far as place and manner of articulation are concerned, the transition is primarily consonantal, but so far as length goes, it is part of the vowel. Even so far as the duration of the transition is concerned, this serves not as a cue to the length of the consonant but to its membership in one or another manner class of such segments (Liberman et al., 1956). If one may draw a conclusion from the published studies concerned with consonant duration (Denes, 1955; Lisker, 1957a, 1958; Liberman et al., 1961a; Fischer-Jørgensen, 1964; Obrecht, 1965), there is a general conviction that consonant length is tied exclusively to the duration of a steadystate acoustic segment.

The discussion so far has been exclusively of the temporal dimension as a segmental property, primarily because it is at this level that difficulties of the kinds mentioned are most apparent, difficulties that make it hard to assess the significance of a good deal of the numerical data available. The literature also includes studies of the temporal properties of entities both smaller and larger than the phonetic segment. Items of the first kind are, for example, the individual periods of the voice fundamental (Lieberman, 1963) and the timing relation between stop articulation and voicing onset (Liberman et al., 1961b; Lisker and Abramson, 1964, 1967, 1970; Abramson and Lisker, 1965, 1970). Speech stretches longer than the segment for which temporal properties have been determined are the syllable, the word, the phrase, and the entire speech piece itself. The durations of these various units are, of course, not independent of one another, and the literature concerned with these larger units is devoted to describing, not their absolute durations, but their durations relative to the durations of units of a different, usually higher, order of magnitude. Thus, for example, Chistovich and her colleagues (1965: 92-95) determined durations for selected syllables in relation to the durations of both the words and sentences in which they were contained, their purpose being to determine regularities from which to derive a model of the temporal organization of speech behavior generally. It might be said that their method of investigation effectively bypassed the segmentation problem, for they did not look for temporal relationships among segment durations, but rather went directly to certain accessible parts of the vocal tract to obtain recordings of the articulatory components involved in the production of their test utterances. They were then able to determine the temporal relations among these components of the articulatory complex. From their work and that of MacNeilage and DeClerk (1969:1233) the conclusion has been drawn that the syllable, or perhaps a "basic syllable" consisting of a consonant and following vowel, is the unit of temporal organization of the articulatory gestures executed in producing what is ordinarily thought to be a sequence of segments. The various articulatory gestures are not adequately defined by membership in a class of segments, for their execution depends significantly on the particular combination of gestures which constitutes the given syllable (Chistovich et al., 1965: 161). Work by these same authors, and by Daniloff and Moll (1968), demonstrates that the articulatory gestures attributed to a particular segment are not in general executed in phase and that the interval of co-occurence is that of a syllable or more. The syllable as a phonetic unit has long been controversial among linguists and owes whatever recognition it enjoys with them to its establishment as a convenience in talking about phoneme distribution. Phoneticians, however, have recognized that certain segmental properties are associated with position within the syllable. Thus Abercrombie (1967:40) states that releasing, or syllable-initial, consonants are of very short duration, while syllable-final, or arresting, consonants are longer, and in presenting durational data phoneticians generally are careful to specify the location within the syllable of the segment in question (e.g., Malécot, 1968). Failure to sort out segments by position within the syllable and in relation to pause can, in fact, lead to statements of dubious value, as for example those of Parmenter and Treviño (1935) on the relative length of voiced and voiceless stops in English.

One kind of evidence advanced in support of the syllable as a unit of temporal organization of speech behavior is that certain acoustic segments said to belong to the same syllable show durations that are negatively correlated (Chistovich et al., 1965:105). Thus, it has been observed repeatedly that in English and some other languages vowels tend to be shorter before voiceless consonants than before voiced and that at the same time the constriction durations of voiceless consonants are greater than those of their voiced counterparts (Laziczius, 1961:120). Nor have convincing counter examples to this generalization been attested. The relation between the durations of vowel and following voiced, as against voiceless, consonant is such that the total durations of the two kinds of sequences tend to approach equality, given a constant overall speaking rate. From data for English trochees produced in isolation (Sharf, 1962:28), it would seem that the combination of a given vowel with either one of a homorganic pair of stops has a duration that is independent of the voicing characteristic of the stop. We might then be prepared to believe that the syllable, or at least the stressed syllable of English, has an inherent duration so far as the vowel and following consonant are concerned, one that varies only with change in overall speaking rate. This, however, is open to question. If we consider total durations of all vowel-consonant combinations, it is certainly not true that these are independent of the particular vowel constituents, nor is it true of the consonants in any respect other than the voicing characteristic of the stop, and possibly also the fricative, consonants. From Sharf's data for English, as well as from Elert's for Swedish (Elert, 1964: 158-163), it appears that the variability in stop closure duration is not sufficient to make up for vowel differences ascribable to their inherent durations (Lehiste, 1970:18-19). Moreover, if we consider vowel-consonant sequences containing consonants of diverse manners of articulation, we find that their components are not so correlated in duration as to yield a fixed syllable duration (Chistovich et al., 1965:111). Thus, given the generalization that low vowels are longer than high ones and that fricatives are longer than stops (Laziczius, 1961:120), it is not the case that a sequence of low vowel and fricative has a duration anywhere near equal to that of a sequence of high vowel and stop consonant. For this reason, Chistovich and her colleagues conclude that if rhythmic patterning of speech activity is to be established, it must be on the basis of temporal regularities at a level higher than the syllable (Chistovich et al., 1965:110).

The observed negative correlation between vowel and succeeding stop consonant that has been reported for a number of languages--e.g., English (Sharf, 1962), French (Belasco, 1958; Delattre, 1962), Russian (Chistovich et al., 1965:99), Norwegian (Fintoft, 1961:35), and Tamil (Lisker, 1958)--has been the object of several kinds of speculative explanations. The generally greater closure durations reported for voiceless stops has most often referred to a feature of relative force of articulation, the fortis-lenis dimension. Thus it is claimed either that voiceless stops and fricatives involve a greater expenditure of articulatory energy and hence longer closures or that fortis obstruents have as a consequence of their more forceful articulation longer closures that are usually voiceless (Jakobson et al., 1952: 26,36,38; Chomsky and Halle, 1968:325). The vowel preceding a fortis stop is shorter than one before a lenis one because the speed of articulation is greater for the former (see, for example, Fujimura, 1961), and consequently the vowel is abbreviated. Alternatively it has been supposed that there is a tendency for vowel-consonant sequences to be produced with a constant total expenditure of effort. If effort is linearly related to duration, then the greater effort required for the fortis stop calls for a reduction in vowel duration (Durand, 1946:172; Belasco, 1958). Still another explanation has been advanced for the greater duration of vowels before voiced stops, namely that, in order to maintain voicing during the articulatory closure, a special adjustment of the larynx must be completed before the closure, and vowel durations of the magnitude found before the voiceless stops do not allow sufficient time for this laryngeal maneuver (Chomsky and Halle, 1968:301). In view of the fact that vowels, in English at least, are as long or longer before other consonants for which no special adjustment of the larynx seems necessary for voicing and that, moreover, the intrinsically long vowels may well be longer before voiceless stops than the intrinsically short ones before the voiced, this last way of accounting for durational differences between allophones of the same vowel phonemes in the two positions is far from convincing. Nor can it be said that attempts to explain durational differences among the consonants of differing manner of articulation merit any more immediate acceptance.

Of speech stretches greater than the syllable, comparatively little information is available until we get to the speech piece as a whole. For certain languages, English for example, it is said that there is a more or less regular alternation of prominent syllables and stretches of one or more syllables of comparatively low audibility. These are the so-called "stress-timed" languages (Abercrombie, 1967:97-98). By contrast, the "syllable-timed" languages presumably exhibit no temporal organization of syllables in larger rhythmic units (Abercrombie, 1967:97). In languages of the first type, the durations of syllables and their component segments that fall between two prominent syllables depend considerably on how many of them there are in the particular piece. The total duration of the piece also appears to determine strongly the durations of component segments and/or syllables. Thus Fónagy and Magdics (1960:192) report that short phrases in Hungarian are produced at a slower rate than are phrases consisting of eight or more segments (op. cit., p. 182). Extensive data for English have been presented by Goldman-Eisler (1954, 1956, 1961, 1967, 1968) to show that an extremely wide range for pieces of less than about 100 syllables, becoming stable for longer pieces at a rate of about 225 syll./min., though with considerable intersubject variation. She determined, however, that the range of variation for short pieces, from about 60 to 600 syll./min., was not a variation in speed of articulation but one in the amount of pausing within the piece (Goldman-Eisler, 1968:26).[9] On the other hand, Laziczius (1961:123) cites durational values for a particular Hungarian vowel in a series of successively longer (in number of syllables) words, and the vowel undergoes progressive shortening. This effect is perhaps in part to be explained as a junctural effect, since proximity to pause seems regularly to be accompanied by an increase in segment duration as part of a reduction in the rate of higher-level units (Lehiste, 1960; Hoard, 1966).

---

[9] It is difficult to believe, however, that uninterrupted pieces are not sometimes judged to have been produced at different speaking rates. There is no certainty, if indeed this is true, as to whether a measure of phonetic segments, syllables, or perhaps words per unit time would best match listeners' judgments of relative rate. Experiments by Osser and Peng (1964), in which phonetic segments per minute were determined for English and Japanese, failed to turn up a significant difference between speakers of the two languages. Since each group judged the other to consist of rapid talkers, the meaning of the negative finding is not clear, nor would a discovered difference be easy to interpret.

The durations of the various units into which the speech piece may be divided are more or less mutually dependent, and consequently the precise significance of absolute time determinations is, in general, subject to question. The nature of the temporal dependencies, both among units of different levels and among coordinate units, is of much greater significance for the establishment of temporal regularities in speech behavior. Exceptions to this are to be found, however, in studies aimed at determining particular time constants of articulation or perception (Lehiste, 1970:6-9). Duration is often, but by no means invariably, to be correlated with the perceptual dimension of temporal length. As an attribute of the acoustic segment, it is sometimes interpreted as length (Abramson, 1962; Obrecht, 1965), sometimes as a cue to consonantal manner of articulation (Gerstman, 1956, 1957; Liberman et al., 1956), force of articulation (Chomsky and Halle, 1968:324), voicing state (Denes, 1955; Lisker, 1957a, 1958; Sharf, 1962), stress (Fry, 1968:370), and perhaps speech rate (Huggins, 1968). At the level of the syllable, little data on duration have been collected, apart from segmental measurements involving the comparison of syllables of differing degrees of prominence. The syllable and the word figure often, however, in discussions of speech rate, the determination of which requires only the decision as to how many of one or the other units are contained in a speech sample and a determination of the total duration of the sample. But with the discussion of speech rate, the focus of interest tends to become the individual speaker and thus peripheral to the central concerns of phonetic research.

REFERENCES

Abercrombie, D. 1964. Syllable quantity and enclitics in English. In In honour of Daniel Jones. 216-222. London: Longmans.

Abercrombie, D. 1967. Elements of general phonetics. Chicago: Aldine.

Abramson, A. S. 1962. The vowels and tones of standard Thai: Acoustical measurements and experiments. Publication 20 of the Indiana University Research Center in Anthropology, Folklore, and Linguistics. Bloomington, Indiana: Indiana University.

Abramson, A. S. and L. Lisker. 1965. Voice onset time in stop consonants: Acoustic analysis and synthesis. 5th Intern. Congr. Acoust., Liège.

Abramson, A. S. and L. Lisker. 1970. Discriminability along the voicing continuum: Cross-language tests. Proc. 6th Intern. Congr. Phon. Sci. 569-573. Prague: Academia Publishing House of the Czechoslovak Academy of Sciences.

Ainsworth, W. A. 1968. Perception of stop consonants in synthetic CV syllables. L & S. 11.139-155.

Avakjan, R. V. 1968. Mimicking the duration of the vowels in aphasia. Z. Phon., Sprachwiss. u. Komm. Fschg. 21.190-193.

Bauernschmidt, A. 1965. Amuzgo syllable dynamics. Lg. 41.471-483.

Belasco, S. 1958. Variations in vowel duration: Phonemically or phonetically conditioned? JAcS. 30.1049-1050.

Bernard, J. R. L-B. 1970. On nucleus component durations. L & S. 13.89-101.

Black, J. W. 1961. Relationships among fundamental frequency, vocal sound pressure, and rate of speaking. L & S. 4.196-199.

Bloomfield, L. 1933. Language. New York: Henry Holt.

Bolinger, D. L. and L. J. Gerstman. 1957. Disjuncture as a cue to constructs. Word 13.246-255.

Bricker, P. D. and S. Pruzansky. 1966. Effects of stimulus content and duration on talker identification. JAcS. 40.1441-1449.

Broadbent, D. E. and P. Ladefoged. 1959. Auditory perception of temporal order. JAcS. 31.1539 (A).

Brooks, M. Z. 1964. On Polish affricates. Word 20.207-210.

Catford, I. C. 1968. The articulatory possibilities of man. Manual of phonetics, ed. by Bertil Malmberg. 309-333. Amsterdam: North-Holland.

Chistovich, L. A., V. A. Kozhevnikov, V. V. Alyakrinskiy, L. V. Bondarko, A. G. Goluzina, Yu. A. Klaas, Yu. I. Kuz'min, D. M. Lisenko, V. V. Lyublinskaya, N. A. Fedorova, V. S. Shuplyakov, and R. M. Shuplyakova. 1965. Rech': Artikulyatsiya i vospriyatiye, ed. by V. A. Kozhevnikov and L. A. Chistovich. Moscow and Leningrad: Nauka. Trans. as Speech: Articulation and perception. Washington: Clearinghouse for Federal Scientific and Technical Information. JPRS. 30.543.

Chomsky, N. and M. Halle. 1968. The sound pattern of English. New York: Harper and Row, Publishers.

Cohen, A., J. F. Schouten and J. 't Hart. 1962. Contribution of the time parameter to the perception of speech. Proc. 4th Intern. Congr. Phon. Sci., Helsinki 1961. 555-560.

Compton, A. J. 1963. Effects of filtering and vocal duration upon the identification of speakers, aurally. JAcS. 35.1748-1752.

Cooper, F. S. 1950. Research on reading machines for the blind. Blindness: Modern approaches to the unseen environment, ed. by Paul A. Zahl. 512-543. Princeton, N. J.: Princeton University Press.

Creelman, C. D. 1962. Human discrimination of auditory duration. JAcS. 34.582-593.

Daniloff, R. and K. Moll. 1968. Coarticulation of lip rounding. J. Speech and Hearing Res. 11.707-721.

Day, R. S. 1970. Temporal order perception of a reversible phoneme cluster. JAcS. 48.95(A).

Delattre, P. 1962. Some factors of vowel duration and their cross-linguistic validity. JAcS. 34.1141-1143.

Delattre, P. 1964. Change as a correlate of the vowel-consonant distinction. Studia Linguistica 18.12-25

Delattre, P. 1968. From acoustic cues to distinctive features. Phonetica 18.198-230.

Delattre, P. and M. Hohenberg. 1968. Duration as a cue to the tense/lax distinction in German unstressed vowels. Intern. Rev. of appl. Ling. VI/4.367-390.

Delattre, P. and M. Monnot. 1968. The role of duration in the identification of French nasal vowels. Intern. Rev. of appl. Ling. VI/3.267-288.

Denes, P. B. 1955. Effect of duration on the perception of voicing. JAcS. 27.761-764.

Denes, P. B. and E. N. Pinson. 1963. The speech chain. Bell Telephone Laboratories, Inc. Baltimore: Waverly Press.

Durand, M. 1946. Voyelles longues et voyelles brèves: Essai sur la nature de la quantité vocalique. Paris: C. Klincksieck.

Eek, A. 1970. Articulation of the Estonian sonorant consonants. I. n and l. Eesti NSV Teaduste Akadeemia Toimetised 19.103-121.

Eek, A. 1970. Some coarticulation effects in Estonian. Academy of Sciences of the Estonian S. S. R. VI.81-85.

Elert, C.-C. 1964. Phonologic studies of quantity in Swedish. Stockholm: Almqvist and Wiksell.

Endres, W.K. 1968. On the experimental evaluation of the interior redundancy in speech. The 6th Intern. Congr. on Acoust., Tokyo. B111-B114.

von Essen, O. 1957. Überlange Vokale und gedehnte Konsonanten des Hochdeutschen. Z. für Phon. u. allgem. Sprachwiss. 10.239-244.

Fairbanks, G., N. Guttman, and M.S. Miron. 1957. Effects of time compression upon the comprehension of connected speech. J. Speech and Hearing Disorders 22.10-19.

Fairbanks, G. and L.W. Hoaglin. 1941. An experimental study of the durational characteristics of the voice during the expression of emotion. Speech Monogr. 8.85-90.

Fairbanks, G. and F. Kodman, Jr. 1957. Word intelligibility as a function of time compression. JAcS. 29.636-641.

Fant, C.G.M. 1960. Acoustic theory of speech production. The Hague: Mouton.

Fant, C.G.M. 1962. Sound spectrography. Proc. 4th Intern. Congr. Phon. Sci., Helsinki 1961. 14-33. The Hague: Mouton.

Fant, C.G.M., J. Liljencrants, V. Maláč, and B. Borovickova. 1970. Perceptual evaluation of coarticulation effects. Speech Transmission Laboratory. Royal Institute of Technology, Stockholm, Sweden. QPSR. 1/1970, 10-13.

Fay, W.H. 1966. Temporal sequence in the perception of speech. The Hague: Mouton.

Fintoft, K. 1961. The duration of some Norwegian speech sounds. Phonetica 7.19-39.

Fischer-Jørgensen, E. 1954. Acoustic analysis of stop consonants. Miscellanea Phonetica 2.42-59.

Fischer-Jørgensen, E. 1963. Beobachtungen über den Zusammenhang zwischen Stimmhaftigkeit und intraoralem Luftdruck. Z. f. Phon., Sprachwiss. u. Kom. Fschg. 16.19-36.

Fischer-Jørgensen, E. 1964. Sound duration and place of articulation. Z. für Sprachwiss. u. Komm. Fschg. 17.175-207.

Fischer-Jørgensen, E. 1969. Voicing, tenseness and aspiration in stop consonants, with special reference to French and Danish. Annual Report III. Inst. Phon. Univ. Copenhagen. 63-114.

Fónagy, I. and J. Fónagy. 1966. Sound pressure level and duration. Phonetica 15.14-21.

Fónagy, I. and K. Magdics. 1960. Speed of utterance in phrases of different lengths. L & S. 4.179-192.

Fry, D.B. 1955. Duration and intensity as physical correlates of linguistic stress. JAcS. 27.765-768.

Fry, D.B. 1968. Prosodic phenomena. Manual of phonetics, ed. by Bertil Malmberg. 365-410. Amsterdam: North-Holland.

Fujimura, O. 1961. Bilabial stop and nasal consonant: A motion picture study and its acoustical implications. J. Speech Hearing Res. 4.233-247.

Gaitenby, J.H. 1965. The elastic word. Status report on speech research, Haskins Laboratories. SR-2. 3.1-3.12.

Gay, T. 1968. Effect of speaking rate on diphthong formant movements. JAcS. 44.1570-1573.

Gay, T. 1970. A perceptual study of American English diphtongs. L & S. 13.65-88.

Gerstman, L.J. 1956. Noise duration as a cue for distinguishing among fricative, affricate and stop consonants. JAcS. 28.160(A).

Gerstman, L. J. 1957. Perceptual dimensions for the fricative portion of certain speech sounds. Doctoral dissertation, New York University.

Gimson, A. C. 1962. An introduction to the pronunciation of English. London: Edward Arnold.

Gleason, H. A., Jr. 1961. An introduction to descriptive linguistics. (revised edition). New York: Holt, Rinehart and Winston.

Goldman-Eisler, F. 1954. On the variability of the speed of talking and its relation to the length of utterances in conversations. British J. of Psych. 45.94-107.

Goldman-Eisler, F. 1956. The determinants of the rate of speech output and their mutual relations. J. of Psychosomatic Res. 1.137-143.

Goldman-Eisler, F. 1961. The significance of changes in the rate of articulation. L & S. 4.171-174.

Goldman-Eisler, F. 1967. Sequential temporal patterns and cognitive processes in speech. L & S. 10.122-132.

Goldman-Eisler, F. 1968. Psycholinguistics: Experiments in spontaneous speech. London and New York: Academic.

Hadding-Koch, K. and A. S. Abramson. 1964. Duration versus spectrum in Swedish vowels: Some perceptual experiments. Studia Linguistica 18. 94-107.

Hanhardt, A. M., D. H. Obrecht, W. R. Babcock, and J. B. Delack. 1965. A spectrographic investigation of the structural status of Ueberlaenge in German vowels. L & S. 8.214-218.

Harrell, R. S. 1958. Some English nasal articulations. Lg. 34.492-493.

Harris, C. M. 1953. A study of the building blocks in speech. JAcS. 25.962-969.

Harris, K. S. 1958. Cues for the discrimination of American English fricatives in spoken syllables. L & S. 1.1-7.

Harris, K. S. Physiological aspects of articulatory behavior. (See this volume.)

't Hart, J. and A. Cohen. 1964. Gating techniques as an aid in speech analysis. L & S. 7.22-39.

Haugen, E. 1956. The syllable in linguistic description. For Roman Jakobson. 213-221. The Hague: Mouton.

Heffner, R-M. S. 1950. General phonetics. Madison: U. of Wisconsin.

Henderson, A., F. Goldman-Eisler, and A. Skarbek. 1966. Sequential temporal patterns in spontaneous speech. L & S. 9.207-216.

Henry, F. 1948. Discrimination of the duration of a sound. J. Exp. Psych. 38.734-743.

Hirsh, I. 1959. Auditory perception of temporal order. JAcS. 31.759-767.

Hoard, J. E. 1966. Juncture and syllable structure in English. Phonetica 15.96-109.

Hockett, C. F. 1955. A manual of phonology. Memoir No. 11 of Intern. J. Am. Ling. Baltimore: Waverly.

Hockett, C. F. 1958. A course in modern linguistics. New York: Macmillan.

Hofmann, T. R. 1967. Initial clusters in English. Quart. Prog. Rep. No. 84., Res. Lab. of Electr., M. I. T. 263-274.

House, A. S. 1961. On vowel duration in English. JAcS. 33.1174-1178.

House, A. S. and G. Fairbanks. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. JAcS. 25. 105-113.

Huggins, A. W. F. 1964. Distortion of the temporal pattern of speech: Interruption and alternation. JAcS. 36.1055-1064.

Huggins, A.W.F. 1968. The perception of timing in natural speech I: Compensation within the syllable. L & S. 11.1-11.

Jakobson, R., C.G.M. Fant, and M. Halle. 1952. Preliminaries to speech analysis: The distinctive features and their correlates. Techn. Rep. No. 13., Acoustics Lab. Cambridge: M.I.T.

Jassem, W., J. Morton, and M. Steffen-Batóg. 1968. The perception of stress in synthetic speech-like stimuli by Polish listeners. Speech analysis and synthesis I. 289-308. Warsaw: Państwowe Wydawnictwo Naukowe.

Jespersen, O. 1932. Lehrbuch der Phonetik. (5th ed.) Leipzig and Berlin: B.G. Teubner.

Johansson, K. 1969. Transitions as place cues for voiced stop consonants: Direction or extent? Studia Linguistica 23.69-82.

John, J.E.J. and J.N. Howarth. 1965. The effect of time distortions on the intelligibility of deaf children's speech. L & S. 8.127-134.

Jones, D. 1950. The phoneme: Its nature and use. Cambridge: W. Heffers & Sons.

Kantner, C.E. and R. West. 1960. Phonetics. (revised ed.) New York: Harper and Brothers.

Kent, R.D. and K.L. Moll. 1969. Vocal-tract characteristics of the stop consonants. JAcS. 46.1549-1555.

Kim, C-W. 1970. A theory of aspiration. Phonetica 21.107-116.

Klumpp, R.C. and J.C. Webster. 1961. Intelligibility of time-compressed speech. JAcS. 33.265-267.

Ladefoged, P. 1964. A phonetic study of West African languages: An auditory-instrumental survey. West African Lg. Monogr. Series, No. 1. Cambridge: Cambridge University.

Ladefoged, P. 1968. Linguistic aspects of respiratory phenomena. Proc. Conf. on Sound Production in Man, Nov. 1966. 141-151. New York: New York Academy of Sciences.

Laziczius, J. 1961. Lehrbuch der Phonetik. Berlin: Akademie Verlag.

Lehiste, I. 1960. An acoustic-phonetic study of internal open juncture. Suppl. to Phonetica. 5.1-54.

Lehiste, I. 1970. Suprasegmentals. Cambridge: M.I.T. Press.

Lehiste, I. and G.E. Peterson. 1961. Transitions, glides and diphthongs. JAcS. 33.268-277.

Lenneberg, E.H. 1967. Biological foundations of language. New York: J. Wiley and Sons.

Liberman, A.M. 1970a. The grammars of speech and language. Cognitive Psychology 1.301-323.

Liberman, A.M. 1970b. Some characteristics of perception in the speech mode. Ch. XVI of Perception and its disorders, res. publ. Assoc. Research Nerv. Mental Dis. XLVIII.

Liberman, A.M., F.S. Cooper, D.P. Shankweiler and M. Studdert-Kennedy. 1967. Perception of the speech code. Psychol. Rev. 74.431-461.

Liberman, A.M., F.S. Cooper, M. Studdert-Kennedy, K.S. Harris, and D.P. Shankweiler. 1968. On the efficiency of speech sounds. Z. für Phon., Sprachwiss. u. Komm. Fschg. 21.21-32.

Liberman, A.M., P. Delattre, L. J. Gerstman, and F.S. Cooper. 1956. Tempo of frequency change as a cue for distinguishing classes of speech sounds. Amer. J. Exp. Psychol. 52.127-137.

Liberman, A.M., K.S. Harris, P. Eimas, L. Lisker, and J. Bastian. 1961a. An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. L & S. 4.175-195.

Liberman, A. M., K. S. Harris, J. A. Kinney, and H. Lane. 1961b. The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. 61.379-388.

Liberman, A. M., F. Ingemann, L. Lisker, P. Delattre, and F. S. Cooper. 1959. Minimal rules for synthesizing speech. JAcS. 31.1490-1499.

Lieberman, P. 1960. Some acoustic correlates of word stress in American English. JAcS. 32. 451-454.

Lieberman, P. 1963. Some acoustic measures of the fundamental periodicity of normal and pathological larynges. JAcS. 35.344-353.

Liiv, G. 1961. Estonian vowels of three degrees of length. Inst. Lg. Lit. Acad. Sci. Estonian S.S.R. 1-8.

Liiv, G. 1962. On the acoustic composition of Estonian vowels of three degrees of length. Eesti NSV Teaduste Akadeemia Toimetised, XI Koide. Uhiskonnateaduste Seeria 3.271-290.

Lindblom, B. E. F. 1963. Spectrographic study of vowel reduction. JAcS. 35.1773-1781.

Lindblom, B. E. F. 1967. Vowel duration and a model of lip mandible coordination. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR. 4.1-29.

Lindblom, B. E. F. and M. Studdert-Kennedy. 1967. On the role of formant transitions in vowel recognition. JAcS. 42.830-843.

Lindner, G. 1969. Einführung in die experimentelle Phonetik. Munchen: Max Hueber Verlag.

Lisker, L. 1957a. Closure duration and the intervocalic voiced-voiceless distinction in English. Lg. 33.42-49.

Lisker, L. 1957b. Linguistic segments, acoustic segments, and synthetic speech. Lg. 33.370-374.

Lisker, L. 1958. The Tamil occlusives: Short vs. long or voiced vs. voiceless? Indian Linguistics, Turner Jubilee I. 294-301.

Lisker, L. (in press). Stop duration and voicing in English. Mélange à la mémoire de Pierre Delattre, ed. by Albert Valdman. The Hague: Mouton.

Lisker, L. and A. S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. Word 20.384-422.

Lisker, L., and A. S. Abramson. 1967. Some effects of context on voice onset time in English stops. L & S. 10.1-28.

Lisker, L. and A. S. Abramson. 1970. The voicing dimension: Some experiments in comparative phonetics. Proc. 6th Intern. Congr. Phon. Sci. 563-567. Prague: Academia Publishing House of the Czechoslovak Academy of Sciences.

Lisker, L., F. S. Cooper, and A. M. Liberman. 1962. The uses of experiment in language description. Word 18.82-106.

Lowe, A. D. and R. A. Campbell. 1965. Temporal discrimination in aphasoid and normal children. J. Speech Hearing Res. 8.313-314.

Lubker, J. F. and P. J. Parris. 1970. Simultaneous measurements of intraoral pressure, force of labial contact, and labial electromyographic activity during production of the stop consonant cognates /p/ and /b/. JAcS. 47.625-633.

Lüdtke, H. 1969. Die Alphabetschrift und das Problem der Lautsegmentierung. Phonetica 20.147-176.

Lüdtke, H. 1970. Sprache als kybernetisches Phänomen. Bibliotheca Phonetica 9.34-50.

MacNeilage, P. F. 1963. Electromyographic and acoustic study of the production of certain final clusters. JAcS. 35.461-463.

MacNeilage, P. F. 1970. Motor control of serial ordering in speech. Psych. Rev. 77.182-196.

MacNeilage, P. F. and J. L. DeClerk. 1969. On the motor control of coarticulation in CVC monosyllables. JAcS. 45.1217-1233.

MacNeilage, P. F., T. P. Rootes, and R. A. Chase. 1967. Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. J. Speech Hearing Res. 10.449-467.

Malécot, A. 1968. The force of articulation of American stops and fricatives as a function of position. Phonetica 18.95-102.

Malécot, A. 1969. The effect of syllabic rate and loudness on the force of articulation of American stops and fricatives. Phonetica 19.205-216.

Malécot, A. 1970. The lenis-fortis opposition: Its physiological parameters. JAcS. 47.1588-1592.

Malécot, A. and P. Lloyd. 1968. The /t/:/d/ distinction in American alveolar flaps. Lingua 19.264-272.

Malmberg, B. 1963. Structural linguistics and human communication. New York: Academic Press.

Martin, J. G. and W. Strange. 1968. The perception of hesitation in spontaneous speech. Perception and Psychophysics 3.427-438.

Mattingly, I. G. 1966. Synthesis by rule of prosodic features. L & S. 9. 1-13.

Mattingly, I. G. and A. M. Liberman. 1970. The speech code and the physiology of language. Information processing in the nervous system, ed. by K. N. Leibovic. 97-117. New York: Springer.

Morton, J. and W. Jassem. 1965. Acoustic correlates of stress. L & S. 8.159-181.

Nasr, R. T. 1960. Phonemic length in Lebanese Arabic. Phonetica 5. 209-211.

Obrecht, D. H. 1965. Three experiments in the perception of geminate consonants in Arabic. L & S. 8.31-41.

Öhman, S. E. G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. JAcS. 39.151-168.

Öhman, S. E. G. 1967. Numerical model of coarticulation. JAcS. 41.310-320.

Osser, H. and F. Peng. 1964. A cross cultural study of speech rate. L & S. 7.120-125.

Parmenter, C. E. and S. N. Treviño. 1935. The length of the sounds of a Middle Westerner. American Speech 10.129-133.

Peterson, G. E. and I. Lehiste. 1960. Duration of syllable nuclei in English. JAcS. 32.693-703.

Peterson, G. E. and J. E. Shoup. 1966. Glossary of terms from the physiological and acoustic phonetic theories. J. Speech Hearing Res. 9.100-120.

Pickett, J. M. 1965. Some acoustic cues for synthesis of the /n,d/ distinction. JAcS. 38.474-477.

Pickett, J. M. and I. Pollack. 1963. Adjacent context and the intelligibility of words excised from fluent speech. JAcS. 35.807(A).

Pike, K. L. 1943. Phonetics. Ann Arbor: University of Michigan Press.

Ruhm, H. B., E. O. Mencke, B. Milburn, W. A. Cooper Jr., and D. E. Rose. 1966. Differential sensitivity to duration of acoustic signals. J. Speech Hearing Res. 9.371-384.

Sawashima, M. Laryngeal research in experimental phonetics (See this volume.)

Schouten, J. F., A. Cohen, and J. 't Hart. 1962. Study of time cues in speech perception. JAcS. 34.517-518.

Schwartz, M. F. 1967. Syllable duration in oral and whispered reading. JAcS. 41.1367-1369.
Scott, R. J. 1967. Time adjustment in speech synthesis. JAcS. 41.60-65.
Shankweiler, D., K. S. Harris, and M. L. Taylor. 1968. Electromyographic studies of articulation in aphasia. Arch. Phys. Med. and Rehabil. 49.1-8.
Sharf, D. J. 1962. Duration of post-stress intervocalic stops and preceding vowels. L & S. 5.26-30.
Sharf, D. J. 1964. Vowel duration in whispered and in normal speech. L & S. 7.89-97.
Singh, S. and J. W. Black. 1965. A study of nonsense syllables spoken by two language groups in varying conditions of sidetone and reading rate. L & S. 8.208-213.
deSivers, F. 1964. A qualitative aspect of distinctive quantity in Estonian. Word 20.28-34.
Slis, I. H. 1970. Articulatory measurements on voiced, voiceless and nasal consonants: A test of a model. Phonetica 21.193-210.
Slis, I. H. and A. Cohen. 1969. On the complex regulating the voiced-voiceless distinction II. L & S. 12.137-155.
Soda, T., Y. Nishida, and H. Suwoya. 1967. Intraoral pressure change in Japanese consonants. Otologia Fukuoka 13, suppl. 1.34-43.
Stetson, R. H. 1951. Motor phonetics. Amsterdam.
Stevens, K. N., A. S. House, and A. P. Paul. 1966. Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. JAcS. 40.123-132.
Sticht, T. G. and B. B. Gray. 1969. The intelligibility of time compressed words as a function of age and hearing loss. J. Speech Hearing Res. 12.443-448.
Studdert-Kennedy, M. and A. M. Liberman. 1963. Psychological considerations in the design of auditory displays for reading machines. Proc. Intern. Congr. on Techn. and Blindness. 289-304.
Subtelny, J. D. and J. H. Worth. 1966. Intraoral pressure and rate of flow during speech. J. Speech Hearing Res. 9.498-518.
Suzuki, H. 1970. Mutually complementary effect of rate and amount of formant transition in distinguishing vowel, semivowel, and stop consonant. QPR 96. Res. Lab. Electr., M.I.T. 164-172.
Thomas, I. B., P. B. Hill, F. S. Carroll, and B. Garcia. 1970. Temporal order in the perception of vowels. JAcS. 48.1010-1013.
Vieregge, W. H. 1969. Untersuchungen zur akustischen Struktur der Plosivlaute. (doct. diss.) Bonn: Rheinische Friederich-Wilhelms-Universität.
Warren, D. W. and S. B. Mackler. 1968. Duration of oral port constriction in normal and cleft palate speech. J. Speech and Hearing Res. 11. 391-401.
Warren, R. M. 1968. Verbal transformation effect and auditory perceptual mechanisms. Psych. Rev. 70.261-270.
Warren, R. M. 1970. Perceptual restoration of missing speech sounds. Science 167.392-393.
Warren, R. M., C. J. Obusek, and R. M.. Farmer. 1969. Auditory sequence: Confusion of patterns other than speech or music. Science 164.586-587.
Westin, K., R. G. Buddenhagen, and D. H. Obrecht. 1966. An experimental analysis of the relative importance of pitch, quantity, and intensity as cues to phonemic distinctions in southern Swedish. L & S. 9.114-126.

Wickelgren, W. A. 1969. Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psych. Rev. 76.1-15.

Yanagihara, N. and C. Hyde. 1966. An aerodynamic study of the articulatory mechanism in the production of bilabial stop consonants. Studia Phonologica IV. 70-80.

Zemlin, W. R., R. G. Daniloff, and T. H. Shriner. 1968. The difficulty of listening to time-compressed speech. J. Speech and Hearing Res. 11.875-881.

Zimmerman, S. A. and S. M. Sapon. 1958. Note on vowel duration seen cross-linguistically. JAcS. 30.152-153.

Zwirner, E. and K. Zwirner. 1966. Grundfragen der Phonometrie. (2nd ed.) Bibliotheca Phonetica. Basel and New York: S. Karger.

A Study of Prosodic Features*

Philip Lieberman
Haskins Laboratories, New Haven+

We will discuss a number of recent advances in the study of the prosodic elements of speech. We will deliberately neglect many recent studies that are based on the unaided senses of a trained observer. We will instead concentrate on studies that would have not been possible in the absence of current techniques for acoustic, physiologic, anatomical, and perceptual experimentation. We shall, moreover, deal with the "linguistic" aspects of intonation. Charles Darwin (1872) noted that cries convey the emotional state of the organism in both man and animals. Darwin was, of course, concerned with the attributes of communication that are common to both man and all other animals, i.e., the nonlinguistic aspects of the suprasegmental prosodic features. The linguistic analysis of the suprasegmental features becomes quite complex inasmuch as these features are also used for the nonlinguistic aspects of speech communication.

We are reserving the term "linguistic" for language-relevant aspects of the suprasegmental features, i.e., those aspects that serve to convey meaning through the medium of language. Linguistic systems differ quite fundamentally from nonlinguistic systems of communication in that individual cries, or phonetic elements, have no inherent meaning. They derive a meaning only after syntactic and semantic analysis. The sound [m] has no inherent "meaning" in a linguistic system. In English, the word man does have a definite meaning. The sound [m] as it is used in producing the word man has, however, no particular meaning. It forms part of a phonetic "coding" of a word which is the "object" that has the linguistic meaning. The sound [m] can also be used to code other words, e.g., am, mama, etc.

The sound [m] can also have a nonlinguistic function. A particular speaker may, for example, use this sound outside the language system to convey a particular emotional state. It might signify that he is happy. The sound [m] as the speaker used it for this nonlinguistic function would have a definite fixed meaning. No syntactic analysis would be necessary to derive its meaning. Note that its meaning would also be idiosyncratic. The listener might or might not know that this sound signified that the particular speaker were happy. The nonlinguistic "meaning" of [m] would not form part of the English language.

The sounds used in human speech thus serve for communication at many levels. We can easily differentiate at least five factors that are transmitted by means of speech:

---

*Chapter prepared for Current Trends in Linguistics, Vol. XII. Thomas A. Sebeok, Ed. (The Hague: Mouton).

+Also, University of Connecticut, Storrs.

1) The speech signal conveys acoustic cues that serve to identify the individual speaker. This aspect of speech communication is quite important. When telephone systems are degraded to the point where they do not transmit these acoustic cues the public begins to complain (Flanagan, 1965). Animal communication also makes use of cries to identify individual animals to their progeny, mates, friends, and associates. Birds, for example, employ such identification signals (Beer, 1969). The acoustic signal also serves to identify the species (Marler and Hamilton, 1966; Greenewalt, 1968) but this aspect of acoustic communication is not relevant to human speech at the present time since Homo sapiens is the only living species that can produce the range of sounds used in speech (Lieberman, 1968b; Lieberman and Crelin, 1971).

The cues that humans use to identify particular speakers involve both the segmental and the suprasegmental phonetic features. Individual speakers indeed may employ different syntactic "styles" that involve different base structures and different optional transformations[1] to convey similar semantic information. It is clear, however, that the suprasegmental features are quite important in establishing the identity of a particular speaker.[2] When the fundamental frequency of a speaker is transposed (by using a Vocoder apparatus) (Flanagan, 1965) it becomes quite difficult to identify the speaker. Transposing the fundamental frequency, of course, changes the perceived pitch of the speaker's voice.

2) The speech signal conveys the linguistic background of the speaker. Individual languages involve language-specific phonetic and syntactic elements. At the phonetic level, language-specific implementation rules (Lieberman, 1970) are involved as well as specific feature ensembles that may be drawn from the set of universal features (Jakobson et al., 1952; Chomsky and Halle, 1968). There are apparently language-specific elements that are manifested in intonation (Ladefoged, 1967; Lieberman, 1967).

3) The speech signal conveys the sex of the speaker. In many languages this occurs at the phonetic level. The fundamental frequency of the speech signal is usually lower for male speakers. This reflects the longer vocal cords that males usually have (Negus, 1949). The vocal cords usually increase in length in males at puberty as the thyroid cartilage grows larger. Adult females also have lower fundamental frequencies than juvenile females since their larynges also grow larger. The great and abrupt increase in the length of the vocal cords is, however, a secondary sexual dimorphism in males. Other acoustic differences also manifest the sex of speakers. Male speakers of English, for example, seem to use lower formant frequencies than do females

---

[1] We will operate within the framework of a generative grammar (Chomsky, 1957, 1968) that makes use of phonemic and phonetic features (Jakobson et al., 1952; Chomsky and Halle, 1968; Postal, 1968).

[2] The fine structure of the fundamental frequency of phonation can even play a part in transmitting the state of health of the speaker. Measurements of the variations in fundamental periodicity, "pitch perturbations," have been used as a diagnostic tool for the early detection of cancer of the larynx as well as other laryngeal pathologies (Lieberman, 1963).

(Peterson and Barney, 1952). These differences may reflect the larger size of male vocal tracts.

4) The speech signal conveys the emotional state of the speaker. Much of the "meaning" of speech is communicated at this level. When we listen to a speaker we may be as aware of the emotional content of the signal, which conveys the speaker's attitude toward the situation, as of the "linguistic" content. In many situations the "linguistic" content of the message, i.e., the part of the message conveyed by the words, is quite secondary. Stereotyped greetings, like Good morning, probably serve as vehicles for emotional information. Stereotyped messages, like the elevator operator's Step to the rear of the car please, may primarily serve as carriers that transmit the emotional state of the speaker.

The "tone" of the speaker's voice may indicate whether he is annoyed at the passengers, whether he is happy, etc. Unfortunately, the information conveyed by the "tone" of the speaker's voice is somewhat ambiguous. The listener really does not know whether an "angry" tone means that the elevator operator is angry at the passengers or that his breakfast was not edible or that his back aches. Emotional information is rarely specific. There are further difficulties insofar as certain emotional nuances are themselves stereotyped. Thus in Newark, New Jersey, and San Francisco, California, different prosodic patterns may signify disdain. The listener must be aware of the speaker's background and current social convention. The "primary" emotional attributes like extreme pain may indeed by stable (Darwin, 1872), but many of the emotional and attitudinal nuances are probably dialect specific. These dialect-specific aspects are in a sense paralinguistic. They are to a certain degree arbitrary. They thus are linguistic in the sense that the relation between "meaning" associated with a sound and the sound is arbitrary. However, they are not like the "linguistic" aspects of the speech signal insofar as there is a direct relation between these signals and their meanings. There is no morphophonemic or syntactic level.

5) The "linguistic" content of the speech signal is naturally of paramount interest to linguists. Certain aspects of the prosodic features convey linguistic information. Like all other phonetic elements, these prosodic features have no meaning in themselves. We shall direct our attention to a review of the state of current research on some of these prosodic features. We will attempt to limit our discussion to the linguistic aspects of speech. This often is difficult when one deals with the prosodic features of intonation, accent, and prominence. Many analyses, e.g., Pike (1945), have attempted to treat what we have termed levels 4 and 5 as an entity. We will, however, attempt to differentiate these aspects of the speech signal, though we recognize that there will always be some uncertainty as to whether a particular prosodic feature is a dialect-specific level 4 or a linguistic level 5 event. Indeed the process by which language develops may, in part, consist of level 4 to level 5 transitions for particular phonetic features, particular words, or syntactic processes. Note that we are not stating that levels 1 to 4 are unimportant.

The Suprasegmental Prosodic Features: Acoustic Correlates

It is convenient and reasonable to consider two elements in connection with the suprasegmental prosodic features. One element is the suprasegmental sentence or phrase intonation, the other element is the class of features like accent and prominence. Whereas the scope of intonation is generally an entire sentence or phrase, the scope of prominence or accent is usually a single syllable, though longer stretches of speech can also be accented or assigned prominence.

The results of many experiments that have involved not only the analysis of speech but speech synthesis in connection with Vocoder equipment[3] have shown that the primary acoustic cue that signals the intonation of an utterance is the fundamental-frequency contour. That is, the manner in which the fundamental-frequency contour varies with respect to time largely determines the intonation of the utterance. Other factors undoubtedly also play a role in defining the perceived intonation of the utterance. The amplitude of the speech signal, for example, generally decreases toward the end of an intonation contour. There also are strong indications that the duration of a syllable is a function of its position within the intonation contour. However, the fundamental frequency at certain points in the intonation contour (Denes, 1959) appears to be the most important acoustic correlate of intonation. We will return to this point again. Many phonetic analyses of intonation have created the impression that the fundamental-frequency contour is the only acoustic correlate of intonation. These analyses[4] assign linguistic significance to minute variations in fundamental frequency throughout the entire sentence. They, in effect, assume that the fundamental-frequency contour must be specified in minute detail throughout the entire intonation contour. Recent experimental evidence, which we will discuss, suggests that this is not the case.

We will begin by discussing one phonetic feature, the breath-group, that describes some of the linguistic functions of intonation (Lieberman, 1967, 1970; Lieberman et al., 1970). We will also discuss other constructs that have been introduced to describe the same linguistic phenomena that the breath-group describes, as well as constructs that are necessary to describe still other linguistic phenomena. Virtually all phonetic and acoustic studies agree that certain acoustic patterns occur in speech that specify intonation patterns. The disagreements and uncertainty are with regard to a) what acoustic parameters are important, b) how these acoustic parameters are generated and controlled by the human speech-production apparatus and c) how many distinct patterns there are and what their linguistic significance is. A description of the acoustic correlates of the breath-group is therefore in order at this point since it will be equivalent at the acoustic level to many of these alternate theoretical constructs. The acoustic description will also hopefully clarify some of the concepts that we shall discuss.

---

[3] The commercial application of Vocoder equipment, which would offer significant economies on high cost circuits like the Atlantic cable, has been delayed for over thirty years by the deficiencies that exist in "pitch extractors" (Flanagan, 1965).

[4] The study presented by Isacenko and Schadlich (1963) is typical of a class of studies that assign linguistic significance to minute (5Hz) variations in fundamental frequency over a long utterance.

In Figure 1 we have reproduced some data (Lieberman, 1967) that shows some of the acoustic parameters associated with a *normal breath-group*. (An equivalent notation is *-breath-group*.) The upper plot in this figure is a quantized sound spectrogram.[5] The darkened areas that are enclosed by "contour lines" represent the relative energy that is present at the frequency plotted on the ordinate scale as a function of time. Time is plotted on the abscissa in seconds. The energy present in a timing pulse is displayed at the two points marked by arrowheads on the abscissa after 0.5 and 1.5 seconds. The speaker uttered the sentence, *Joe ate his soup*. Note, for example, the energy concentration at approximately 200 Hz (a Hz is equivalent to a cycle per second) at t = 0.4 seconds which is the spectrogram's representation of the first formant of the vowel of the word *Joe*. It is possible to determine relative energy levels by means of the quantized spectrogram.

The second plot from the top in Figure 1 is the smoothed fundamental frequency of phonation as a function of time. The fundamental frequency was derived by measuring the tenth harmonic on a narrow bandwidth spectrogram.

The uppermost plots in Figure 1 thus show that the fundamental frequency of phonation and sound energy both decrease at the end of the *-breath-group*. The duration of segmental phonemes also appears to increase at the end of the breath-group. (Note the duration of the closure interval of the stop /p/, which is longer than the closure interval of the stop /t/.) Also note that the relative energy balance changes at the end of the breath-group. (There is less energy in the higher formants of the vowel of *soup* as the breath-group ends.)

These observations will become more meaningful as we discuss the results of current research on the behavior of the larynx. We will also discuss the significance of the two lower plots in Figure 1. Similar acoustic correlates of what we have termed the *normal breath-group* have been reported by Chiba (1935) in an early study which made use of electronic analysis equipment.

Recent studies by Jassem (1959), Hadding-Koch (1961), Revtova (1965), Fromkin and Ohala (1968), Matsui et al. (1968), Ohala (1970), Vanderslice (1970), and Lieberman et al. (1970) also have derived similar acoustic correlates. Armstrong and Ward (1926) and Jones (1932) also postulate similar acoustic correlates for Tune I. The Jones and Armstrong and Ward studies, of course, were conducted without the benefit of modern instrumentation. Some of the acoustic correlates of the normal breath-group may also be seen in Jones (1909) and Cowan (1936) where perceived pitch and fundamental frequency, respectively, are plotted as functions of time for relatively large

---

[5] The wide-band filter of the sound spectrographs that are usually used for the analysis of speech has a bandwidth of 300 Hz. This bandwidth accepts at least two harmonics of the glottal source for typical male speakers. The wide-band filter thus will manifest the formant frequencies rather than the individual harmonics (Flanagan, 1965). The formant frequencies are uniquely determined by the cross-sectional area function of the supralaryngeal vocal tract.

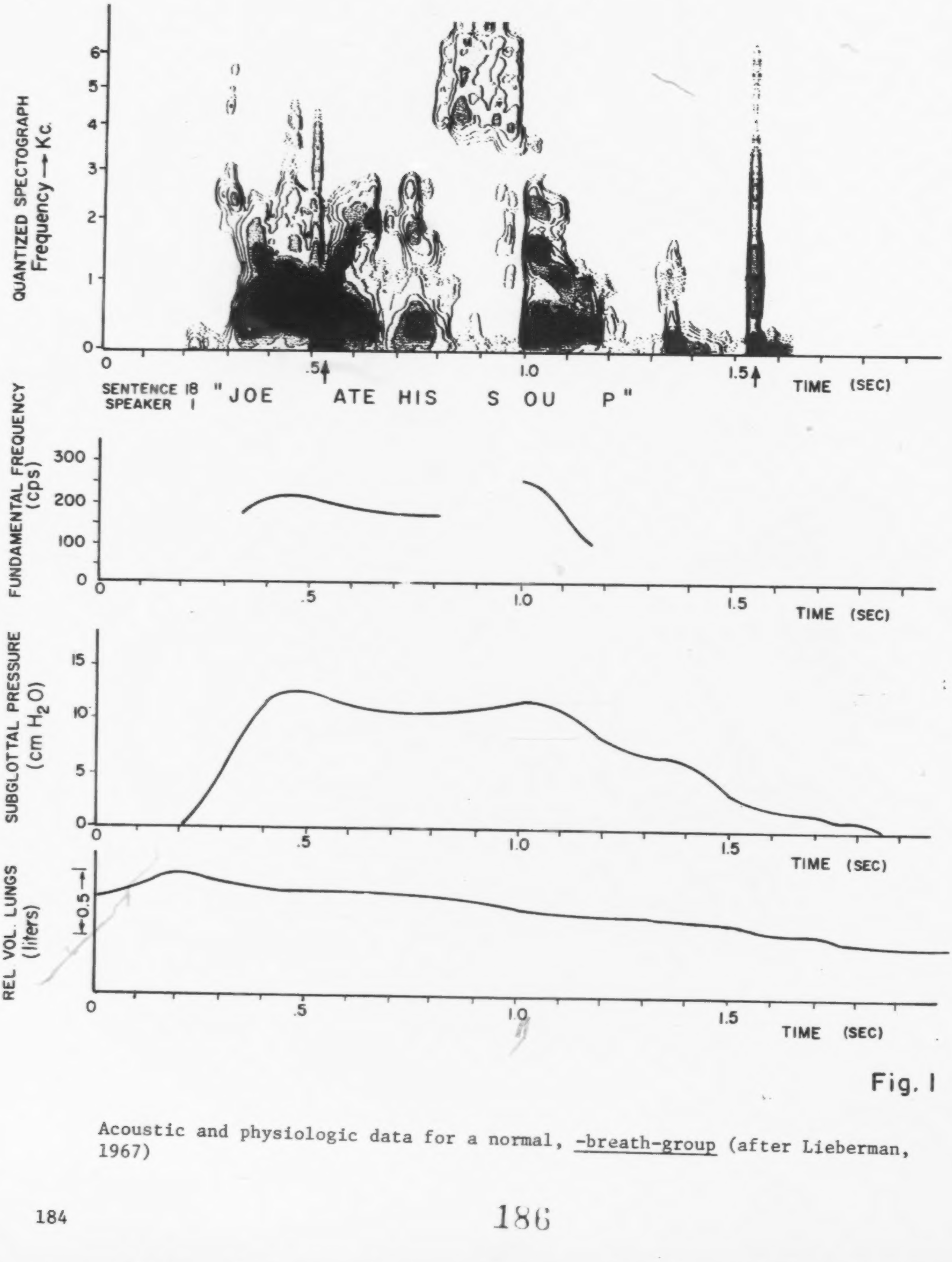


Fig. I

Acoustic and physiologic data for a normal, -breath-group (after Lieberman, 1967) data samples. Pike (1945) also postulates similar acoustic correlates for the "final pause [\\]." The Trager and Smith (1951) "terminal juncture [#]" also appears to have similar acoustic correlates (Hadding-Koch, 1961; Lieberman, 1965).

Note that the acoustic correlates that we have discussed all are relative measurements over a span of time that encompasses a string of segmental phonetic elements. The normal breath-group is thus a true suprasegmental. The "pause" notation developed by Pike (1945) and Trager and Smith (1951) also is an implicit suprasegmental since the amplitude and fundamental frequency of the pause are relative measures that are defined with respect to the pitch-amplitude contour that precedes the terminal. The contour that precedes the terminal thus is an intimate part of the terminal contour. The Jones (1932) and Armstrong and Ward (1926) Tune I notation, of course, is an explicit suprasegmental phonetic entity. Harris (1944) also postulates, without any experimental evidence, suprasegmental intonation "morphemes."

In Figure 2 we have reproduced some data that shows some of the acoustic parameters associated with a +breath-group (Lieberman, 1967). Note that the primary difference between the fundamental frequency contour of this utterance and Figure 1 is that the fundamental frequency rises at the end of the breath-group. In some instances, a +breath-group ends with a level fundamental-frequency contour. The significant point is that it does not end with a falling fundamental-frequency contour. Similar acoustic and psychoacoustic data is again available (Chiba, 1935; Jassem, 1959; Fromkin and Ohala, 1968; Mattingly, 1966, 1968; Matsui et al., 1968; Ohala, 1970; Vanderslice, 1970; Lieberman et al., 1970). Armstrong and Ward (1926) and Jones (1932) also postulate similar acoustic correlates for Tune II as does Pike (1945) for the "tentative pause [\]." Trager and Smith (1951) postulate similar acoustic correlates for the "terminal junctures [\] and [\\]." They differentiate between the juncture [\] which ends with a level pitch contour and [\\] which ends with a rising fundamental-frequency contour. Note that these intonation contours are also true suprasegmentals (whether the transcription is in terms of Tune II, terminal juncture [\\], etc.). The fundamental frequency at the end of the contour is defined with respect to its behavior earlier in the contour. The listener must keep track of the entire contour in order to "decode" the final fundamental-frequency contour. Note that this automatically makes intonation contours into speech segmenting devices that have fairly long spans.

In contrast to the "long span" suprasegmental intonation contours other prosodic features have shorter spans. The feature which we shall call prominence (Lieberman, 1967, 1970; Lieberman et al., 1970) generally spans only a single syllable. Its acoustic correlates involve local increases in the fundamental frequency of phonation, the amplitude of the speech signal, and the duration of the segment (Fry, 1958; Jassem, 1959; Lieberman, 1960, 1967, 1970; Wang, 1962; Morton and Jassem, 1965; Fonagy, 1966; Ladefoged, 1969; Lehiste, 1961; Rigault, 1962; Hadding-Koch, 1961; Bolinger, 1958). The phonetic quality of the prominent syllable also may change (Lehiste and Peterson, 1959; Fry, 1965). The formant frequencies of the +prominent syllable show less coarticulation with adjacent segmental phonetic elements (Lindblom, 1963).

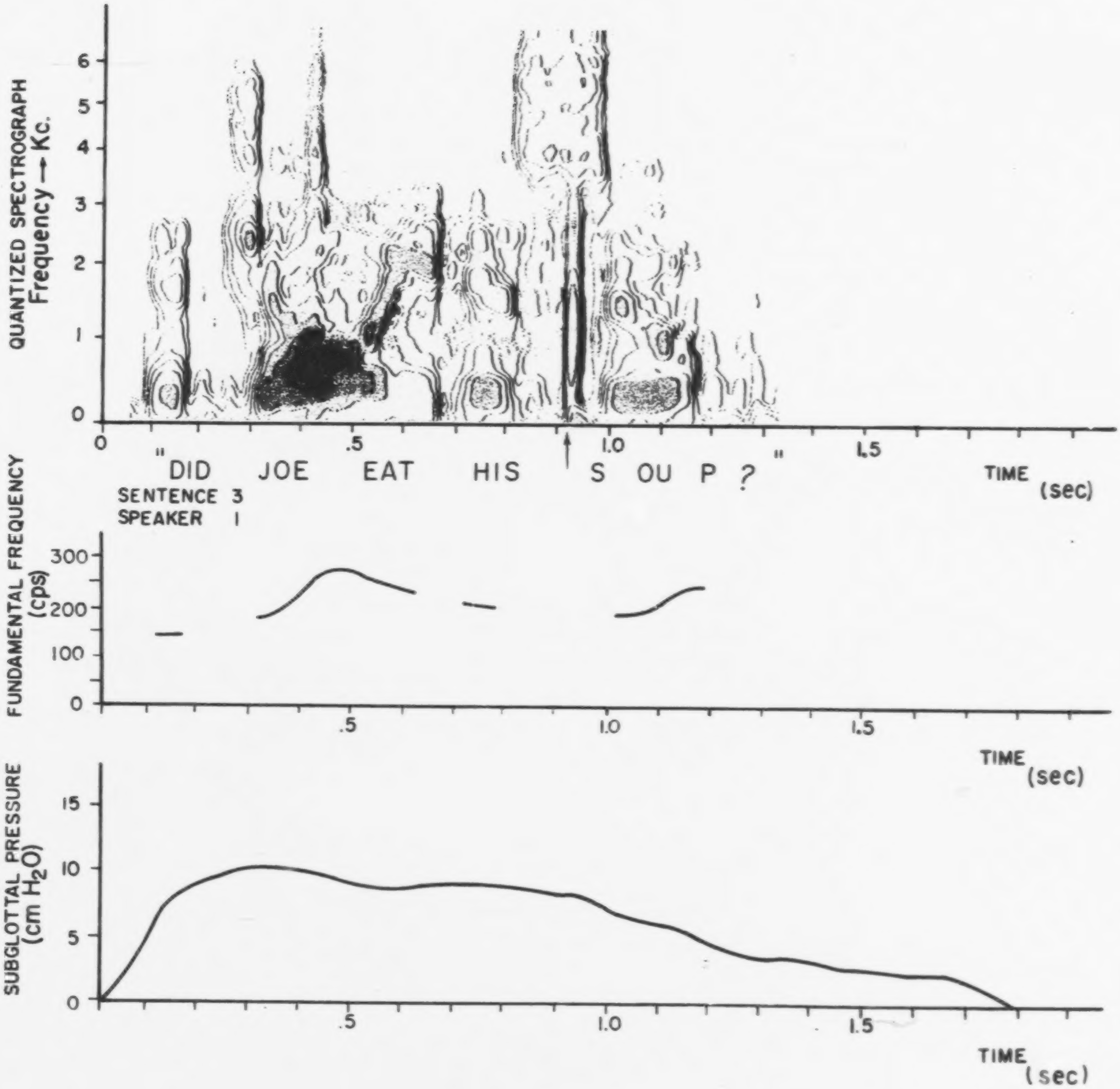


Fig. 2

Acoustic and physiologic data for a marked, +breath-group (after Lieberman, 1967)

The acoustic correlates of prominence all seem to reflect increased activity of the muscles that are involved in speech production (Fonagy, 1966; Harris et al., 1968). All or some of the acoustic correlates can be used to mark a prominent syllable (Fry, 1958; Jassem, 1959; Lieberman, 1960). Note that the acoustic correlates that manifest prominence (excepting changes in formant frequencies) are again "prosodic" effects that interact with the long-span intonation to effect the total prosodic structure of an utterance.

Another phenomenon that appears to be a short-span prosodic feature is what we shall term accent. Whereas prominence appears to involve the synergetic activity of many muscles which act in concert to produce increases in the relative fundamental frequency, amplitude, and duration of a segment, an additional feature of accent exists. Accent appears to involve only contrasts in the fundamental-frequency contour. A syllable marked with the feature accent thus may have the acoustic correlate of a sudden decrease in fundamental frequency (Bolinger, 1958; Morton and Jassem, 1965; Katwijk and Govaert, 1967; Öhman, 1968; Barron, 1968; Vanderslice, 1970). We have deliberately avoided using the term "stress" which appears in many studies of the prosodic features since stress appears to be an abstract linguistic construct (Chomsky and Halle, 1966).[6] In certain instances linguistically stressed syllables are not manifested phonetically by either prominence or accent (Lieberman, 1965; Barron, 1968). The feature of accent is also not always necessarily used to manifest underlying linguistic stress. The segmental "tone" systems that occur in so many languages may involve the feature (or features) of accent at the phonetic level (Öhman, 1968; Wang, 1967). It is important to continually bear in mind the dichotomy between the phonetic and semantic levels of language. Phonetic features in themselves have no inherent "meaning." The status of the term "stress" in generative linguistic studies, which unfortunately conflicts with its use in many phonetic studies, motivates our terminology.

Studies of the Larynx and Subglottal Vocal Mechanism

We have briefly discussed the acoustic correlates of some of the prosodic features that appear to have a reasonable basis in quantitative experimental evidence. There are other possibilities which need to be explored and we shall return to this topic. It is, however, necessary to first review the status of recent research on the anatomical mechanisms that appear to generate the acoustic correlates of the prosodic features. In order to understand the range of possible prosodic features we need to understand the constraints of the speech-production mechanism. Recent studies have fortunately made new information available that should be of benefit for further research on the prosodic features.

---

[6] There is no general agreement on terminology. Vanderslice (1970), for example, for similar reasons also avoids using the term "stress" at the phonetic level. He uses the term accent in about the same way as we do but he uses the term "emphasis" for the feature prominence, "cadence" for the acoustic correlates of the unmarked, -breath-group, and the term "endglide" for the acoustic correlates of the marked, +breath-group. Vanderslice's choice of different terminology appears to be, in part, motivated by theoretical differences concerning the articulatory maneuvers that underlie the acoustic correlates of these prosodic features as well as the status of the motor theory of perception. We will come to these aspects of modern theory later in our discussion. While theoretical differences exist, there appears to be, however, general agreement concerning the relevant acoustic phenomena.

The acoustic theory of speech production shows that the acoustic speech signal can be regarded as the product of a source and a filter function (Chiba and Kajiyama, 1958; Fant, 1960). For voiced sounds, the source is the quasiperiodic series of "puffs" of air that exit from the larynx as the vocal cords rapidly move together and apart. The filter function is determined by the area function of the supralaryngeal vocal tract. As a first approximation, the glottal source and the supralaryngeal vocal tract do not interact. There are some interactions which we will discuss, but we can differentiate between the controlled variations of the supralaryngeal vocal tract's filter function and the controlled changes of the glottal source.

The fundamental frequency of phonation, which is the primary physical correlate of perceived pitch (Flanagan, 1965), is determined by the rate at which the vocal cords adduct and abduct. The energy content and amplitude of the glottal source, i.e., the glottal volume velocity waveform, which excites the supralaryngeal vocal tract is, in turn, determined by the rate at which the vocal cords move. When the vocal cords move faster and more abruptly, the glottal waveform more closely approximates a "pulse" (Timcke et al., 1958). The energy in the higher portions of the glottal spectrum is enhanced under these conditions (Flanagan, 1965). Since the fundamental frequency of a typical male speaker is about 120 Hz whereas the first and second formant frequencies of a vowel like /a/ are 700 and 1100 Hz (Fant, 1960), increasing the "high-frequency" energy content of the glottal source will increase the amplitude of the speech signal (Fant, 1960). Changes in fundamental frequency and amplitude are thus largely determined by the larynx for voiced sounds. It therefore is essential to know how the larynx functions in order to construct a viable phonetic theory.

The Myoelastic-Aerodynamic Theory of Phonation

In the early nineteenth century Johannes Müller (1848) first developed the concepts that have resulted in the myoelastic-aerodynamic theory of phonation (Van den Berg, 1958). This theory, which accounts for the known behavior of the larynx, states that the vocal cords (or vocal folds) rapidly move together and apart during phonation as a result of aerodynamic and aerostatic forces. The vocal cords, in other words, passively move inwards and outwards as forces developed by the airstream rapidly alternate. The laryngeal muscles can adjust the initial position of the vocal cords, which determines whether phonation will or will not take place. The laryngeal muscles can also adjust the tension and mass of the vocal cords (Hirano & Ohala, 1969), which influences the manner in which the vocal cords move. The motive force for phonation is, however, provided by the airstream out from the lungs. The air pressure of the air in the lungs, which during phonation is nearly identical to the subglottal air pressure (Mead and Agostini, 1964), determines, in part, the rate at which the vocal cords move. Subglottal air pressure thus is an important factor in determining the fundamental frequency of phonation. The subglottal air pressure is itself a function of both the impedance of the glottis (i.e., the resistance which the larynx offers to the airflow) and the force generated by the subglottal respiratory system. Since the glottal impedance is relatively high during phonation (Van den Berg, 1957) the activity of the subglottal respiratory system enters as a factor in the articulatory maneuvers that underlie the acoustic correlates of the prosodic features. The relative importance of the laryngeal muscles and the subglottal respiratory system to the control of fundamental frequency is a crucial factor in determining the articulatory implementation of the prosodic features. The investigation of this problem has shown that the larynx is a rather complex device. We will attempt to review some of the pertinent studies without becoming too involved in the physiology and anatomy of the larynx and the subglottal respiratory system.[7]

Air Pressure and Fundamental Frequency

The question which we shall review is beguilingly simple. An observer notes a change in the fundamental frequency of phonation during the production of a sustained note while a speaker is singing, or within the fundamental-frequency contour associated with a short utterance. The observer wants to know whether the change in f_0 (fundamental frequency) follows from a change in the subglottal air pressure developed by the subglottal respiratory system or whether the f_0 change follows from a change in the tension of the muscles of the larynx. The answer to this question appears to be that the larynx has many "modes" of phonation and that the effects of changes in the activity of the subglottal or laryngeal muscles on f_0 are different in different modes of phonation.

The first quantitative experiments on the dynamics of the larynx involved the excitation of excised larynges in which the "lateral" tension on the vocal cords (Van den Berg, 1960) was changed by simulating the activity of the cricothyroid muscle (Müller, 1848). Müller in these experiments was able to simulate the activity of this muscle by an arrangement of pulleys, strings, and weights. He was able to change the force that this muscle would exert on the vocal cords while he simulated the flow of air out from the lungs by blowing air through a tube beneath the excised larynx. Müller found that three conditions had to be satisfied in order for phonation to take place. The vocal cords had to be moved inward from the open position that they assume for respiration and a minimum laryngeal tension and a minimum subglottal air pressure were necessary. Different combination of laryngeal tension and subglottal air pressure resulted in different fundamental frequencies of phonation. Müller could not precisely measure either the fundamental frequency of phonation (he subjectively matched pitch) or the subglottal air pressure, but modern refinements of this experiment have replicated his results.

---

[7] The reader can refer to Ladefoged et al. (1958), Mead and Agostini (1964), and Bouhuys et al. (1966) for detailed discussions of the mechanics of the subglottal system as well as techniques for the measurement of subglottal air pressure (Lieberman, 1968b). The myoelastic-aerodynamic theory of phonation was challenged by Husson (1950), who proposed that the muscles of the larynx provided the motive force for phonation. Husson claimed that the laryngeal muscles actively contracted in order to produce each fundamental period. The role played by subglottal pressure in determining the fundamental frequency of phonation was therefore minimal in Husson's erroneous theory (Van den Berg, 1958).

Van den Berg (1957, 1960) in a series of experiments with excised human larynges has found that there are two distinct "registers" in which phonation occurs. The laryngeal muscles, by adjusting the position and shape of the vocal cords, determine the register. In the chest register the vocal cords have a thick cross section, the part of the vocal cords that is set into motion is long, the tension applied by the cricothyroid muscle is relatively low, some "medial compression" applied by muscles like the interarytenoids and lateral cricoarytenoids is present, and the vocal cords collide in an "inelastic manner" as they move together in each fundamental period (Negus, 1949; Timcke et al., 1958; Van den Berg, 1957, 1960, 1968; Lieberman, 1967; Hirano and Ohala, 1969). In the "falsetto" register the vocal cords have a thin cross section, high lateral forces, a small vibrating mass, and elastic collisions, or no collisions, as the vocal cords move together in each fundamental period.

The vocal cords tend to move more abruptly in the chest register since their thick cross section (Hollien and Curtin, 1960) allows a nonlinear force, the Bernouilli force, to be realized as the air moving out from the lungs is forced through the glottal constriction (Van den Berg, 1957; Flanagan, 1965; Lieberman, 1967). The glottal excitation therefore has more high-frequency energy for phonation in the chest register. The sensitivity of the larynx to changes in subglottal air pressure also varies for the two registers. Van den Berg (1960) in experiments with excised human larynges found that the sensitivity of the larynx to changes in subglottal air pressure ranged from 2.5 Hz to 20 Hz/cm H_2O. In these experiments the tension and configuration of the vocal cords of an excised larynx were held constant while the subglottal air pressure was changed. The tensions and positions of the vocal cords were then set to another set of parameters and the subglottal air pressure was again varied. In this manner a number of plots relating fundamental frequency and subglottal air pressure were obtained. The only criticism that can be leveled at experiments of this sort is whether the air pressures, tension, and positions that the experimenter imposes on the vocal cords of the excised larynges are typical of those that occur *in vivo*. Experiments in which speakers are asked to sing notes while the subglottal air pressure is artificially and abruptly changed indicate that Van den Berg's results are probably realistic.

Lieberman, Knudson, and Mead (1969), for example, performed an experiment in which a single speaker attempted to sing sustained notes while sinusoidal modulations in air pressure were imposed on his subglottal air pressure. The singer sang at both "loud" and "soft" levels at different pitch levels. The rate of change of fundamental frequency was found to vary between 2 and 20 Hz/cm H_2O when the fundamental-frequency variations of the sustained notes were correlated with the sinusoidal air-pressure modulations. The speaker in this experiment did not attempt to match any reference tones as he sang.

A number of experiments have been conducted where a speaker sings a note while his chest is abruptly pushed inward by a light push. This causes his subglottal air pressure to abruptly rise. Low rates of change of f_0 with respect to air pressure (from 2.0 to 5.0 Hz/cm H_2O) have been reported by Ladefoged (1962), Öhman (1968), and Fromkin and Ohala (1968) using this technique. In a recent experiment (Ohala and Ladefoged, 1970) somewhat higher rates of change of f_0 with respect to subglottal air pressure variation (to 10 Hz/cm H_2O) have been reported using this technique. Other studies have correlated fundamental-frequency changes with subglottal air-pressure variations during the production of short sentences. Lieberman (1967) obtained a value of about 17 Hz/cm H_2O by this technique. The activity of the laryngeal muscles was, however, not monitored in this study and some of the variations in f_0 that were ascribed to subglottal air-pressure variations may have been due to the activity of laryngeal muscles tensing. An analysis of the data of Fromkin and Ohala (1968), in which the activity of several laryngeal muscles was monitored, however, shows that the rate of change of f_0 with respect to subglottal air-pressure variation is 12.5 Hz/cm H_2O (Lieberman et al., 1970). It is interesting to note that this same speaker apparently showed less sensitivity to variations in subglottal air pressure (about 5 Hz/cm H_2O) when his activity was monitored while he sang.

Some of the differences between the rates of change of f_0 with respect to subglottal air pressure that we have discussed may be perhaps ascribed to experimental artifacts that arise in the measurement of subglottal air pressure. The physiology of the respiratory system can make the measurement of subglottal air pressure from measurements derived from esophageal balloons (Ladefoged, 1969) rather involved (Bouhuys et al., 1966; Lieberman, 1968a). It nonetheless is clear that the sensitivity of the larynx to changes in subglottal air pressure is variable. The theoretical laryngeal model developed by Flanagan and Landgraf (1968) and Flanagan (1968) predicts that the variation of f_0 with subglottal air pressure will be different for different "modes" and different "registers" of phonation. This model indicates that the larynx is least sensitive to variations in air pressure for phonation in the chest register that involves inelastic collision of the vocal cords. The larynx is most sensitive to air pressure for falsetto register phonation with elastic collision. The total range of variation that the model predicts is 2 to 20 Hz/cm H_2O.

The experimental data on excised larynges as well as data derived from both sustained "singing" and speech in humans thus supports this laryngeal model. We can tentatively conclude that the larynx can either be sensitive or insensitive to variations in subglottal air pressure. The laryngeal configurations used by trained singers probably result in minimum sensitivity since this would simplify the pitch-control problem. The larynx assumes a different posture in singing (Sundberg, 1969). In singing, controlled variations in f_0 probably are the consequence of laryngeal maneuvers. In speech production, speakers apparently may use laryngeal configurations that result in a relatively high sensitivity of f_0 to air-pressure variations (Lieberman, 1967; Lieberman et al., 1970; Kumar and Ojamaa, 1970).

Although it would be far "simpler" if all changes in f_0 during speech were exclusively due to laryngeal maneuvers (Vanderslice, 1967, 1970; Fromkin and Ohala, 1968; Öhman, 1968; Ohala, 1970), this does not appear to be so. The data reported by Fromkin and Ohala (1968), which forms the substantive base of the claim that laryngeal maneuvers exclusively generate the controlled f_0 changes of intonation, indeed indicates that both subglottal pressure changes and laryngeal maneuvers must be taken into account (Lieberman et al., 1970).

"Uncontrolled" f_0 Variations

The theoretical model of the larynx that we have noted (Flanagan and Landgraf, 1968; Flanagan, 1968) also explains other aspects of f_0 variation during speech. Peterson and Barney (1952) in their study of vowel formant frequencies note that different vowels appear to have slightly higher or lower average fundamental frequencies. Similar results have since been noted by House and Fairbanks (1953), Lehiste and Peterson (1959), and Swigart and Takefuta (1968). Vowels that have low first-formant frequencies, e.g., /u/ and /i/, have higher fundamental frequencies. Vowels that have high first-formant frequencies, e.g., /a/, have lower fundamental frequencies. The Flanagan and Landgraf (1968) model of the larynx indicates that these effects, which involve offsets of about 10 Hz for an average f_0 of 120 Hz, are due to aerodynamic coupling between the supralaryngeal vocal tract and the larynx. The presumed independence of the glottal source and the supralaryngeal vocal tract is, as we noted, only a first approximation.

The model developed by Flanagan (1968) also indicates that the sensitivity of the larynx to changes in subglottal air pressure will be affected by the supralaryngeal vocal-tract configuration. The highest rates of change of f_0 with respect to subglottal air pressure occur for vowels with low first-formant frequencies. The model predicts that the vowel /a/ which has a high first-formant frequency will result in the lowest variations in f_0 with changes in subglottal air pressure, all other parameters being equal. These variations may, in part, account for some of the different values for the sensitivity of f_0 to subglottal air-pressure variation that have been obtained for sung vowels, where /a/ is usually produced, and speech.

We have been oversimplifying our discussion of the relationship between subglottal air pressure and fundamental frequency. The transglottal air pressure is actually the factor that we should keep in mind when we discuss these variations (Flanagan, 1965; Lieberman, 1967). We have implicitly assumed that the supraglottal air pressure stays constant while the subglottal air pressure varies. This is, of course, true in studies where a speaker sings a single, sustained vowel. In the production of connected speech this is, however, not the case, as the oral air pressure abruptly builds up during the production of stops like /b/ where the lips close, as well as, to a lesser degree, for other consonantal sounds. These variations in transglottal air pressure will cause variations in f_0 that are concomitant with segmental phonetic elements (Öhman, 1968; Lieberman et al., 1970).

Still other variations in f_0 will arise from coarticulation phenomena (Lindblom, 1963). The larynx is continually reset from its closed phonation position to more open positions for the production of unvoiced segmental phonemes. These transitions are comparatively slow. It takes almost 100 msec to move the vocal cords from an unvoiced to a voiced configuration (Lieberman, 1967; Lisker et al., 1969; Sawashima et al., ms.). Variations in fundamental frequency are thus to be expected as the vocal cords either finish a closing maneuver or begin to anticipate an opening maneuver.

These perturbations of the f_0 contour can be correlated with voiced and unvoiced stops (House and Fairbanks, 1953; Ojamaa et al., 1970). They may be secondary cues for the perception of voiced and unvoiced stops (Haggard, 1969).

The larynx is obviously not an isolated appendage of the human body. It is quite possible that gross skeletal maneuvers can affect the configuration and the tension of the vocal cords through the complex system of ligament, cartilage, and muscle that connects the larynx with the skeletal frame (Sonninen, 1956). Some of the data that traditionally has been cited as supporting evidence for the mechanical interactions affecting f_0 should perhaps be reappraised in the light of the aerodynamic interactions that can occur between the larynx and the supralaryngeal vocal tract. The higher f_0's associated with /u/ and /i/ are sometimes interpreted as evidence for mechanical interaction between the muscles of the tongue and the larynx (Ohala and Ladefoged, 1970). The higher f_0 of these vowels, however, appears to be an aerodynamic effect. Radiographic data (Perkell, 1969) of the vowels makes it difficult to see why both of these vowels should have higher f_0's from mechanical interactions while the vowel /a/ has a lower f_0.

Ohala and Hirano (1967) and Ohala and Hirose (1969) have obtained electromyographic data that shows that the sternohyoid muscle is active when a speaker sings at either high or extremely low fundamental frequencies. Whether similar mechanisms play a role during connected speech is still a question. Öhman (1968) has suggested that such maneuvers may underlie a proposed phonetic feature of -accent which results in an f_0 fall. Ohala (1970) presents data that shows increased sternohyoid activity accompanying falling f_0 contours, but these f_0 changes could also be the consequence of either falling subglottal air pressure or of the opening of the larynx in anticipation of an unvoiced stop.

Some of the distinctions that have been occasionally made between subglottal air pressure and laryngeal maneuvers are meaningless. Subglottal air pressure is the consequence of both the impedance offered by the larynx when the vocal cords are in their adducted or nearly adducted phonation position (Lieberman, 1967) and the activity of the subglottal respiratory system which forces air out from the lungs. Before an unvoiced segment, subglottal air pressure will fall. This will affect the f_0 contour (House and Fairbanks, 1953; Ojamaa et al., 1970). The activity of the larynx that causes this effect is, however, an articulatory manifestation of the segmental phonetic feature -voiced (Chomsky and Halle, 1968) rather than an articulatory manifestation of a prosodic feature. In other instances (Ladefoged et al., 1958; Ladefoged, 1962, 1968, 1969; Lieberman et al., 1970), changes in subglottal air pressure can be observed that clearly are the consequence of maneuvers of the subglottal respiratory system.

Phonetic Theories

We have presented a brief review of some of the factors that cause uncontrolled variations in f_0. These phenomena perhaps explain why many phoneticians have been reluctant to work with "objective," electronically derived, fundamental-frequency contours. These contours frequently contain many errors since it is extremely difficult to derive fundamental frequency by means of electronic devices (Flanagan, 1965). However, even when the electronic devices work, they show minute variations in fundamental frequency that are not acoustic correlates of prosodic features. The human observer will not pay any attention to these variations in the framework of the prosodic feature system. He may perhaps use some of these variations as secondary cues for the perception of the segmental phonetic elements that generate them (Lehiste and Peterson, 1959; Haggard, 1969). Some of the variations in f_0 that occur in speech may simply be the result of chance variations in laryngeal muscle tension or the activity of the subglottal respiratory system. Speakers do not appear to take great care in producing exactly the same f_0 contour for the same utterance (Lieberman, 1967). When the f_0 contours of the "same" sentence are compared for several speakers, startling differences can be seen (Lieberman, 1965, 1967; Rabiner et al., 1969). The utterances, of course, may not be the "same" at the emotional, "level 4" aspect of communication. Contours that have rather different "fine structures" with respect to their f_0 variations do not appear to have any linguistic import. Listeners are unable to differentiate the contours at any linguistic level though they may ascribe different emotional contexts to the contours (Lieberman and Michaels, 1962).

Phonetic theories like that of Isacenko and Schadlich (1963) are unconvincing since they rely on small 5 to 10 Hz variations over long intonation contours in synthesized signals. Linguistic studies like Bierwisch (1966) that attempt to derive "grammatical rules" that will specify intonation contours in this detail are thus overspecified. Preliminary perceptual experiments with synthesized speech appear to show that only a few gross factors are important when the f_0 contour that accompanies an utterance is specified; the direction of the terminal contour (whether it falls or rises) and the point of the major prominence (if any occurs) of the utterance must be specified. Listeners will accept almost all other variations in the f_0 contour. They are, however, sensitive to the duration of each segmental phonetic element with respect to its position in the utterance (Mattingly, 1966, 1968; Matsui et al., 1968). Some phenomena that have traditionally been associated with f_0 in the detailed transcriptions like those of Kingdon (1958), Schubiger (1958), and Halliday (1967) may perhaps have durational correlates.

The perception of intonation by linguists often appears to be "contaminated" by their analysis of other levels of language. In an experiment on the perception of intonation (Lieberman, 1965) linguists trained in the Trager-Smith (1951) notation were asked to transcribe a set of eight utterances. Electronic processing equipment was available that abstracted the acoustic correlates of the prosodic features that the linguists were ostensibly transcribing from the words of the message. When the linguists were presented with these isolated prosodic contours, which modulated a fixed vowel, they were unable to transcribe the pitch levels and junctures that they noted on hearing the complete utterance. The linguists' prosodic transcriptions were, moreover, more accurate when they listened to the isolated acoustic features of fundamental frequency and acoustic amplitude as functions of time. The pitch levels and junctures of the Trager-Smith notation apparently depended on the linguistic information conveyed by the words of the message. The Trager-Smith notation appears to be a device whereby semantic differences carried by different underlying phrase markers can be recorded (Lieberman, 1967). The linguist using this system "hears" the pitch levels that transcribe the "pitch morpheme" that he wishes to present. The notation of "pitch morphemes" follows from the "attitudinal meanings" that Pike (1945) states are conveyed by the prosodic features. Pike is essentially discussing some of the nonlinguistic, "level 4" functions of the prosodic features. Trager and Smith invent pitch morphemes that convey the semantic information that they are unable to account for with the immediate constituent grammar that is the basis of their linguistic theory (Postal, 1964).

The Motor Theory of Speech Perception

One of the most influential developments of recent years is the modern version of the motor theory of speech perception (Liberman et al., 1967). The motor theory essentially states that the constraints imposed by the human speech production are structured into a "speech perception mode." The motor theory thus also states that speech is perceived in a different manner than other acoustic signals. There is a large body of experimental evidence that supports this theory though many points still are in dispute. It is important to note that the motor theory does not state that listeners must "learn" to decode speech through an explicit knowledge of the process of speech production. It thus is irrelevant that listeners who are unable to speak can perceive speech (Lenneberg, 1967). There are undoubtedly special neural detectors in man that are structured in terms of the sounds that the human speech apparatus makes. Similar neural devices have been found in frog (Capranica, 1965) and in cat (Whitfield, 1969). The general motor theory is reviewed in this volume by M. Studdert-Kennedy. It is of interest here since it accounts for some otherwise perplexing phenomena in the perception of the prosodic features.

Several studies have noted that the stressed syllables of English bisyllabic words like *rebel* and *rebel* may receive the phonetic feature that we have called *+prominence* when they occur in isolation.[8] One of the principal articulatory maneuvers that underlies the acoustic correlates of *+prominence* is a momentary increase in subglottal air pressure. The listener's responses to the stressed syllables indicate that he appears to be responding to the magnitude of the subglottal air-pressure peak (Ladefoged, 1962, 1968, 1969; Ladefoged and McKinney, 1963; Lehiste and Peterson, 1959; Fonagy, 1966; Lieberman, 1967). Jones (1932) is perhaps the modern source for this theory for the perception of +prominence. (Jones uses the term stress where we use prominence.) The motor theory of speech perception is most illuminating when we consider the interactions that can occur between +prominent syllables and the sentence intonation's f_0 contour. In Figure 3, two fundamental frequency contours are schematized. Both contours were used with a standard carrier phrase in a controlled psychoacoustic experiment (Hadding-Koch and Studdert-Kennedy, 1964). The listeners in this experiment said that both of these contours were questions that had the same terminal rise. Note that the two contours in Figure 3 have different terminal f_0 rises. The listeners were not merely responding to the physically present fundamental-frequency signal. They instead appeared to be evaluating the terminal fundamental-frequency contours in terms of the underlying articulatory maneuvers that would be present in natural speech.

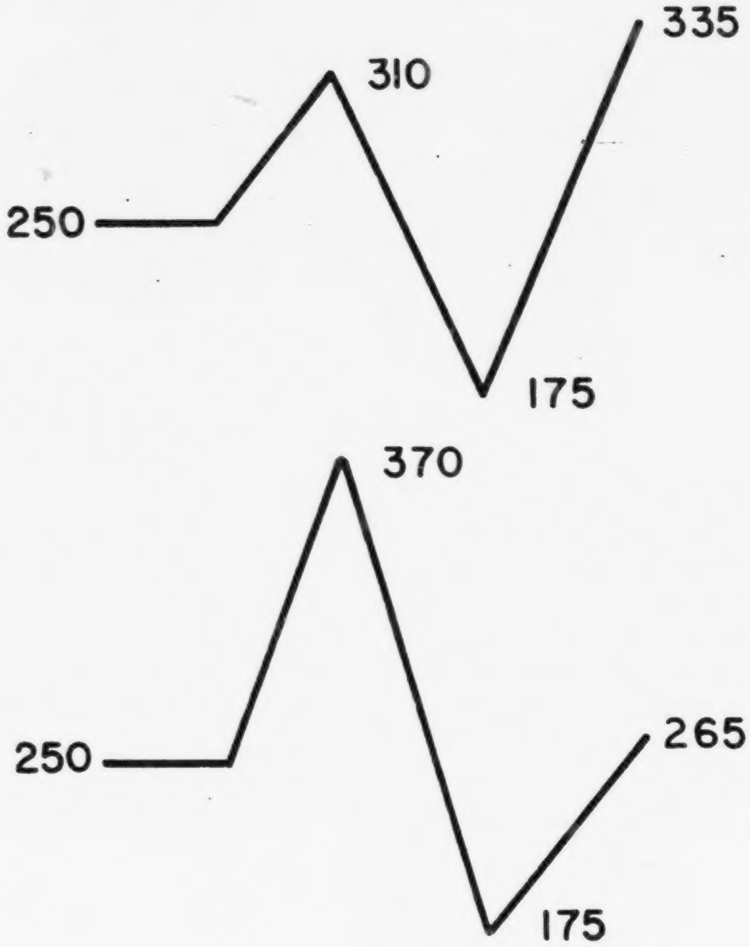
---

[8] When lists of isolated words that have contrasting stress patterns are read in phonetic experiments, the readers will use the feature *+prominence* to manifest primary stress. The stress levels generated by the rules of the phonologic component of the grammar (Chomsky and Halle, 1968) do not, however, appear to be manifested by the feature *prominence* in normal discourse. It is necessary to construct an utterance in which semantic information is carried by a "morpheme" of "emphasis" or "focus" in the underlying phrase marker (Chomsky, 1968; Postal, 1968). These morphemes may be ultimately realized phonetically in different ways. The ultimate result of certain optional transformations and phonologic rules may, for example, yield the sentence, *Tom did run home*. Other transformations and phonologic rules could yield, *Tom RAN home*. (where the capitals indicate the presence of *+prominence* at the phonetic level). In other words, the occurrence of the feature *+prominence* usually seems to be attributable to the presence of some morpheme in the underlying phonologic component of the grammar.

The data presented by Lieberman (1967), Fromkin and Ohala (1968), Ohala (1970), and Lieberman et al. (1970) all show that yes-no questions in English are produced with +breath-groups where the tension of certain laryngeal muscles increases at the end of the breath-group to effect a "not falling" f_0 contour. The -breath-group, which is used for short unemphasized declarative sentences in English, ends with a falling terminal f_0 contour that is a consequence of the falling subglottal air pressure that occurs at the breath-group's end (Lieberman, 1967; Fromkin and Ohala, 1968; Ohala 1970; Lieberman et al., 1970). The data in Figures 1 and 2 show examples of a -breath-group and a +breath-group. Note that the +breath-group also ends with a falling subglottal pressure function. The increased tension of the laryngeal muscles counters the falling subglottal air pressure to produce a rising terminal f_0 contour in Figure 2.

We have noted that the articulatory maneuvers that underlie the prosodic feature +prominence include a momentary increase in subglottal air pressure. The momentary increase in subglottal air pressure will produce an increase in the fundamental frequency since the fundamental frequency of phonation is a function of both laryngeal tension and transglottal air pressure.

The listeners in the Hadding-Koch and Studdert-Kennedy experiment appear to be evaluating the intonation contours in Figure 3 in terms of their "knowledge" of the articulatory bases of these features. They "know" that rising terminal f_0 contours must be the consequence of increased laryngeal tension. They "know" that the nonterminal f_0 peak is a consequence of a peak in subglottal air pressure. They also appear to "know" another fact about speech production that we have not yet discussed. Nonterminal peaks in the subglottal air pressure function will result in a lower subglottal air pressure after they occur. This effect can be seen in the data of Lieberman (1967) as well as in the independent data of Fromkin and Ohala (1968) which is discussed in Lieberman et al. (1970). This "air-pressure perturbation" effect appears to be a consequence of the physiology of the lungs. The elastic recoil of the lungs and the respiratory muscles which regulate subglottal pressure (Mead and Agostini, 1964) appear to be "programmed" in terms of the overall breath-group, whereas independent instructions result in the pressure peak associated with +prominence. The +prominence pressure peak lowers the volume of the lungs which, in turn, lowers the air pressure generated by the elastic recoil (Bouhuys et al., 1966; Lieberman, 1967:54,71,98-100; Lieberman et al., 1970; Kumar and Ojamaa, 1970).


335
310
250
175
370
250
265
175

Synthesized intonation contours that listeners perceived to have the "same" terminal pitch contours (from data of Hadding-Koch and Studdert-Kennedy, 1964)

Fig. 3

The "articulatory decoding" that the listeners use in perceiving the terminal rises in Figure 3 thus goes as follows: the lower contour's +prominence f_0 peak is 370 Hz, whereas the upper contour's is 310 Hz; since a higher f_0 implies the presence of a greater subglottal air pressure peak, the listener knows that the lower contour employed a greater nonterminal subglottal air pressure peak; the listener now "knows" that the terminal subglottal air pressure of the lower contour is lower than the upper contour's. This follows from the air pressure perturbation effect. An equivalent degree of laryngeal tension would thus result in a lower terminal fundamental frequency. The listener thus evaluates the terminal fundamental-frequency contours in terms of the laryngeal tension that would be present in natural speech.

It is important to remember that this "motor theory perception" of intonation does not imply that the listeners actually "know" at a conscious level any of the relationships that we have discussed. People "know" many complex relationships at some neural level without any conscious knowledge of the fact. The act of respiration itself involves many "reflexes" and "regulatory mechanisms" that are not part of the conscious knowledge of the casual "breather." We all must breath but few of us are aware of the maneuvers that we must employ to generate a relatively steady subglottal air pressure over the nonterminal portions of an utterance (c.f., Figs. 1 and 2). We nonetheless "know" how to regulate the variable pressure that would be generated by the elastic recoil of our lungs (Mead and Agostini, 1964; Ladefoged et al., 1958). The "knowledge" of these aspects of speech production probably are the result of a long evolutionary process in man. Modern man's hominid ancestors gradually acquired the anatomical mechanisms that are involved in human speech (Lieberman and Crelin, 1971). The neural decoding mechanism or neural detectors that are involved in the decoding of speech are thus part of man's evolutionary endowment, and we should not really be surprised to find relatively complex pattern detectors. Similar detectors have been found in frogs and cats (Capranica, 1965; Whitfield, 1969). Birds probably employ similar ones (Greenewalt, 1968), and we are presumably not less well endowed.

The point of view that we have explicated, i.e., the structure that is imposed on the production and the perception of speech by the anatomical and physiologic constraints of the human speaker, also is consistent with the intonational forms that human languages tend to use most. If we think of the simplest way in which a speaker can regulate subglottal air pressure to speak, we arrive at the intonation pattern that is most common in the languages of the world (Chiba, 1935; Hadding-Koch, 1961; Revtova, 1965; Lieberman, 1967). This pattern, which involves a terminal falling f_0 contour, follows from the simplest, or "archtypal," pattern of muscular control that can be used to generate a relatively steady subglottal air-pressure contour over the course of an utterance. The simplest pattern obviously involves a state of minimal laryngeal control over the course of the utterance. At the end of the utterance, the fundamental frequency of phonation will rapidly fall since the subglottal air pressure must change from a positive to a negative air pressure. A negative air pressure is the natural consequence of the requirements of respiration. The speaker has to get air into his lungs in order to breathe. The vegetative process of respiration thus determines the form of what we have called the unmarked, or normal, breath-group

(Fig. 1). All people and all languages obviously do not use the simplest "unmarked" state (Chomsky and Halle, 1968). The +breath-group, which is "marked," obviously involves a more complex pattern of muscular activity where the laryngeal muscles must be tensioned at the end of the breath-group in order to counter the falling f_0 contour that would result from the "vegetative" fall in subglottal pressure.

Different languages and different speakers also probably do not make use of the "archtypal," unmarked breath-group even when they retain the falling terminal f_0 contour for most utterances. Idiosyncratic and language-specific patterns of laryngeal activity are often used to produce f_0 variations throughout the nonterminal portions of the breath-group (Lieberman, 1967; Fromkin and Ohala, 1968; Lieberman et al., 1970; Kumar and Ojamaa, 1970). The situation is no different from the implementation of other phonetic features. Idiosyncratic and language-specific modifications also occur (Lieberman, 1970).

We have noted that a number of studies (Bolinger, 1958; Hadding-Koch, 1961; Fromkin and Ohala, 1968; Morton and Jassem, 1965; Barron, 1968; Vanderslice, 1967, 1970; Ohala, 1970) indicate that linguistic stress can be manifested by sudden falls in f_0. These sudden falls in f_0 could be implemented either by tensing laryngeal muscles that would cause fundamental frequency to fall or by relaxing laryngeal muscles that raise f_0 when they tense. There is no a priori reason to assume that all phonetic features must be implemented by tensing a particular muscle. Phonetic features do not stand in a one-to-one mapping with particular muscles. Even when there is a close relation between a particular phonetic feature and a muscle, e.g., nasality and the levator palatini muscle which controls the velopharyngeal port, the implementation of the feature may involve relaxing the muscle in question (Lieberman, 1970). It is therefore possible that abrupt falls in f_0 could be implemented by relaxing muscles that in their tensed state maintain a higher f_0. Either the cricothyroid muscle which applies lateral tension to the vocal cords or the muscles that adduct the vocal cords and apply medial compression (Van den Berg, 1960) could do this.

Several studies (Vanderslice, 1967; Fromkin and Ohala, 1968; Ohala, 1970) have stated that these abrupt falls in f_0 are the consequence of tensing the sternohyoid muscle. This muscle can alter the vertical position of the larynx and a lower laryngeal position is said to produce a lower fundamental frequency (Vanderslice, 1967). The data in these studies show that the nonterminal falling f_0 contours are sometimes the consequence of some laryngeal maneuver. (Falling f_0 is seen where the subglottal pressure is steady.) The activity of the sternohyoid muscle, however, does not appear to be related to these f_0 falls. The sternohyoid muscle has been shown to be active in the maneuvers that singers use to sing at both low and high fundamental frequencies (Hirano and Ohala, 1969; Ohala and Hirose, 1969). The data of Ohala and Hirose (1969) however show that "any gesture of speech or nonspeech that would most likely require a lowering or fixation of the hyoid bone tended to show an increase in the activity of this muscle." The position of the larynx and hyoid bone do change during connected speech. This motion, however, appears to be an articulatory maneuver that is directed at lowering the formant frequencies of segmental phonetic elements (Perkell, 1969) as well as the sustension of phonation in +voiced stops like /b/. The total volume of the supralaryngeal vocal tract expands during the closure interval of these stops in order to sustain voicing (Perkell, 1969). The larynx also may be lowered at the end of phonation as part of the general adjustment of the larynx for inspiration. When the vertical position of the larynx is correlated with the average fundamental frequencies of the vowels (Peterson and Barney, 1952; Swigart and Takefuta, 1968; Perkell, 1969; Sundberg, 1969) it is apparent that f_0 is not correlated with larynx height. The highest and lowest larynx heights, /i/ and /u/, have the highest average fundamental frequencies. The average fundamental frequencies of these vowels are apparently due to aerodynamic coupling with the larynx (Flanagan and Landgraf, 1968) rather than larynx height adjustment.

The emphasis on showing that tensing a laryngeal muscle will lower f_0 in the Vanderslice (1967), Fromkin and Ohala (1968), and Ohala (1970) studies follows from their belief that the subglottal air pressure contour plays virtually no part in determining the f_0 contour. The falling f_0 contour that occurs at the end of a -breath-group thus must have a laryngeal correlate. Since there is no evidence in the electromyographic data of Fromkin and Ohala (1968) that the laryngeal muscles that maintain f_0 relax, the presence of a laryngeal muscle that lowers f_0 as it is tensed is postulated. The close correlation that is manifested between the f_0 and subglottal air-pressure contours for the -breath-groups (the unemphasized declarative sentences) in the data of Fromkin and Ohala (1968) and Ohala (1970) is felt by these authors to be fortuitous. Ohala and Ladefoged (1970) recently have, however, found that air-pressure variations can change fundamental frequency at rates as high as 10 Hz/cm H_2O, whereas they earlier believed that the highest rate that occurred in speech was about 3 Hz/cm H_2O. The controversy regarding the sensitivity of the fundamental frequency to subglottal air-pressure variations, as we noted earlier, apparently reflects the fundamental character of the laryngeal source. The fundamental frequency of phonation can be insensitive to air-pressure variations as it is in the "modes" of phonation that appear to be used in singing. If the sensitivity of the larynx is measured in this mode of phonation, air pressure will have a negligible effect. The larynx, however, can also phonate in "modes" that make it sensitive to air-pressure variations. The data of Fromkin and Ohala (1968) and Ohala (1970), despite the claims made by these authors, show that the larynx is sensitive to air-pressure variations during the production of speech at a rate of about 12.5 Hz/cm H_2O.

Öhman (1968) in a study of the accent system of Swedish has accounted for a set of rather complex variations by means of two prosodic features which we shall call accent down and accent up. These features appear to be sufficient to generate the falling f_0 contours that manifest phonetic stress in English (Morton and Jassem, 1965; Barron, 1968) as well as some of the phenomena discussed by Vanderslice (1967, 1970), Fromkin and Ohala (1968), and Ohala (1970). (These features would, however, not be necessary to account for the falling terminal f_0 contour of the archetypal normal breath-group (-breath-group) which follows from the falling subglottal air-pressure contour.[9]) Bolinger (1958) described a range of phenomena in

---

[9] In some instances, adult speakers can be observed who start to move their larynx towards its open respiratory configuration before the end of a -breath-group. The laryngeal muscles will, as they abduct the vocal cords,

English that also can be accounted for by means of these features. We are following Bolinger's precedent in our terminology. The implementation of +accent up would involve a laryngeal maneuver that raised f_0, whereas the implementation of +accent down would involve a laryngeal maneuver that lowered f_0. Note that in contrast to the feature +prominence, which can involve an increase in laryngeal tension that raises f_0 (Fromkin and Ohala, 1968; Harris et al., 1968; Lieberman et al., 1970; Lieberman, 1970; Ohala, 1970) the feature +accent up would only involve activity in the larynx. Its articulatory implementation would thus be more localized than the feature +prominence which appears to involve heightened muscular activity throughout the vocal tract (Fonagy, 1966; Harris et al., 1968).

The two features accent up and accent down may perhaps be the phonetic manifestations of the segmental "tones" that have been noted in many languages (Chang, 1958; Abramson, 1962; Wang, 1957). These tones clearly interact with the f_0 contour of the sentence, and they appear to be independent of +prominence (Chang, 1958). The intonational transcriptions developed by Jassem (1952), Lee (1956), Kingdon (1958), and Halliday (1967), which are based on their perception of "meaningful" prosodic events, may also reflect these two features of accent.

The studies of Bolinger (1958, 1961), Vanderslice and Pierson (1967), Nash (1967), and Crystal (1969), as well as those of Kingdon, Halliday, Schubiger, and Hadding-Koch, which we have cited, indicate that other prosodic features that are imposed on the breath-group also must be investigated. Armstrong and Ward (1926) and Jones (1932) in their perceptually based studies, for example, note the presence of an element of emphasis which may extend over an entire breath-group (a "tune" in their notation). It is not clear at this time whether this involves an additional feature or whether the scope of the feature +prominence can be extended over an entire breath-group. In like manner, Bolinger (1958, 1961) establishes a convincing case for features that result in gradual changes of f_0 over a comparatively long part of a breath-group. It is again not clear whether

---

lower f_0 by increasing the vibrating mass, lowering the tension of the vocal cords, and lowering the subglottal air pressure by reducing the glottal impedance (Lieberman, 1967). These premature opening maneuvers probably should be regarded as idiosyncratic variations on the archetypal form of the -breath-group (Lieberman, 1970) rather than as implementations of the feature +accent down. They are in the same class as the variations in voicing onset observed in the production of stops by Lisker and Abramson (1964) for isolated individual speakers of English. Other speakers, e.g., Speaker 1 in Lieberman (1967) and the speaker in Fromkin and Ohala (1968), do not open their larynges until the end of phonation in a -breath-group. If we classify idiosyncratic implementations of a feature as manifestations of other features, we would have to conclude that individual speakers had phonetic components that made use of different feature complexes.

Note that a premature opening of the larynx near the end of a -breath-group merely emphasizes the falling f_0 contour that usually is the consequence of the falling subglottal air pressure. The articulatory gesture of opening the larynx is not in opposition to the falling subglottal air pressure. It enhances the trend already established. Phonetic features must act distinctively.

these "ramps" should be regarded as the result of many small steps of *+accent down* or *+accent up* (depending on the direction of the f_0 contour) or whether new features should be introduced.[10] The electromyographic techniques that are described by K. S. Harris in this volume will undoubtedly prove useful in resolving these and other questions. The introduction of physiologic and acoustic techniques has provided a reasonable advance over the theory presented by Stetson (1951). The introduction of new or refined techniques that allow the activity of muscles to be correlated with articulatory, physiologic, and acoustic data promises similar advances in the near future.

Multileveled Versus Binary Features

A problem that is perhaps more susceptible to controlled psychoacoustic experiments is whether prosodic features are multivalued or binary. This problem is, of course, not simply confined to the prosodic phonetic features. It has been raised many times with regard to the segmental phonetic features. The question can be partially resolved in terms of the mechanism available to the phonetic component of a grammatical theory. If phonetic features stand in a close relation to the acoustic signal (Jakobson et al., 1952) or to individual muscles or muscle groups (Ladefoged, 1967; Chomsky and Halle, 1968; Fromkin, 1968), multivalued features will have to be introduced.

If phonetic features are instead regarded as "state functions" at the articulatory level rather than as specific, invariant, muscular commands or articulatory maneuvers, a fairly powerful phonetic component must be introduced into the grammar to yield the actual articulatory maneuvers that underlie speech (Lieberman, 1970). A state function is, for example, the feature *consonantal* which does not involve a particular muscle or articulatory maneuver. The phonetic component of the grammar will involve "implementation rules" that can generate multivalued acoustic or articulatory phenomena from an input ensemble of binary phonetic features. It therefore is possible to have multivalued phenomena like the stress levels postulated by Trager and Smith (1951) even though binary phonetic features form the input to the phonetic component. The phenomenon that Bolinger and Gerstman (1957) and Lieberman (1967) termed "disjuncture" together with the features of *prominence*, *accent down* and *accent up* could, in theory, combine to map out a multivalued stress system. Disjuncture has the acoustic correlate of an unfilled pause. It therefore could combine with any or all of the other prosodic features that we have discussed to provide a distinct physical basis for multivalued stress levels.

Perceptual studies by Bolinger and Gerstman (1957), Lieberman (1965), and Barron (1968) suggest, however, that human listeners cannot differentiate more than two levels of stress in connected speech. Hadding-Koch (1961) attempts to correlate perceived stress levels in the Trager-Smith system with acoustic measurements of Swedish discourse, but the results are not conclusive. In certain restricted contexts (Hart and van Katwijk, 1969) multivalued stress decisions can be made by trained listeners. These effects, as Hart and van Katwijk point out, may be experimental artifacts. The particular choice of features that is used to map out the stress levels does not appear to be a factor in these experiments, and a tentative conclusion is that only two levels of stress can be mapped out by the prosodic features. The process of vowel reduction which appears to be a consequence of the stress-assignment rules in English (Chomsky and Halle, 1968) can provide a physical basis for a third level of stress. It is important to remember that we are using the term "stress" here to signify the linguistically determined "level" that the rules of the phonologic component assign to a vowel. We are not using the term to signify a phonetic feature. A listener thus can perceive the stress levels of an utterance by means of internal computations that involve his knowledge of the phonologic stress assignment rules of his language and the constituent structure of the utterance. There need be no acoustic or phonetic events that specifically map out the stress levels (Lieberman, 1965; Barron, 1968).

---

[10] The author is inclined to speculate that the case of emphasis extending over an entire breath-group can be treated as a special case of *+prominence* where the scope is the entire breath-group. C. J. Bailey (pers. comm.) has developed some convincing grammatical arguments for this approach. The gradual contours described by Bolinger, however, would appear to involve additional prosodic features.

In closing we must take note of one of our prefatory remarks. We have deliberately neglected many studies of intonation that are based on the unaided senses of a trained observer who transcribes what appear to be meaningful prosodic events. We have instead concentrated on recent studies that are based on quantitative acoustic, anatomical, and physiologic data as well as psychoacoustic data. This does not imply that we believe that auditorily based studies are useless. They bring to light many phenomena that would be overlooked in the small data ensembles that generally form the basis of more "quantitative" studies. We have also taken the liberty of making some arbitrary definitions concerning particular phonetic features that may not always agree with the definitions of these terms in other studies. Some degree of arbitrariness is, however, necessary in this regard in light of the differences in theory and method that differentiate various studies. The reader can, it is hoped, make the necessary name changes if he finds them desirable. The situation has hitherto been rather anarchic and we have been forced to bring some degree of order that may sometimes appear arbitrary by redefining several terms. The reader also will note that we have not ventured into the area that we termed "level 4" functions of prosody, e.g., Halliday (1967) and Crystal (1969). As linguistic theory develops, many of these functions will doubtlessly be amenable to linguistic analysis, but, as Crystal himself (1969a) notes, "before any attempt to integrate intonation with the rest of a description is likely to succeed, various preliminaries have to be gone through....One cannot assume that everyone means the same thing by such labels as 'confirmatory.'" We therefore have attempted to deal only with some of the prosodic features that appear to have a clear linguistic function and that have been investigated by means of quantitative procedures.

REFERENCES

Abramson, A. S. 1962. The vowels and tones of standard Thai: Acoustical measurements and experiments. Bloomington: Indiana University, Research Center in Anthropology, Folklore, and Linguistics, #20.

Armstrong, L. E. and I. C. Ward. 1926. Handbook of English intonation. Leipzig and Berlin: B. G. Teubner.

Barron, M.E. 1968. Relation of acoustic correlates to linguistic stress for several English words. Quar. Progress Rept. of Research Lab. Elec., M.I.T. Cambridge, Mass.: M.I.T.

Beer, C.G. 1969. Laughing gull chicks: Recoginition of their parents' voices. Science 166.1030-32.

Bierwisch, M. 1966. Regeln für die Intonation deutscher Sätze. Studia Grammatica 7.99-201.

Bolinger, D.L. 1958. A theory of pitch accent in English. Word 14.109-49.

Bolinger, D.L. 1961. Generality, gradience; and the all or nones. The Hague: Mouton.

Bolinger, D.L. and L.J. Gerstman. 1957. Disjuncture as a cue to constructs. JAcS. 29.778.

Bouhuys, A., D.F. Proctor, and J. Mead. 1966. Kinetic aspects of singing. J. Appl. Physiol. 21.483-96.

Capranica, R.R. 1965. The evoked vocal response of the bullfrog. Cambridge, Mass.: M.I.T. Press.

Chang, N.T. 1958. Tone and intonation in the Chengtu dialect (Szechuan, China). Phonetica 2.59-85.

Chiba, T. 1935. A study of accent, research into its nature and scope in the light of experimental phonetics. Tokyo: Phonetic Society of Japan.

Chiba, T. and M. Kajiyama. 1958. The vowel, its nature and structure. Tokyo: Phonetic Society of Japan.

Chomsky, N. 1957. Syntactic structures. The Hague: Mouton.

Chomsky, N. 1968. Aspects of the theory of syntax. Cambridge, Mass.: M.I.T. Press.

Chomsky, N. and M. Halle. 1968. The sound pattern of English. New York: Harper and Row.

Cowan, M. 1936. Pitch and intensity characteristics of stage speech. Archives of Speech, Supplement I.1-92.

Crystal, D. 1969. Prosodic systems and intonation in English. Cambridge: Cambridge Univ. Press.

Crystal, D. 1969a. Review of Intonation and grammar in British English by M.A.K. Halliday. Language 45.378-93.

Darwin, C. 1872. The expression of emotion in man and animals. London: J. Murray.

Denes, P. 1959. A preliminary investigation of certain aspects of intonation. L & S. 2.106-22.

Fant, C.G.M. 1960. Acoustic theory of speech production. The Hague: Mouton.

Flanagan, J.L. 1965. Speech analysis, synthesis and perception. Berlin: Springer-Verlag.

Flanagan, J.L. 1968. Studies of a vocal-cord model using an interactive laboratory computer. Preprints of the Kyoto Speech Symposium, August 1968. Kyoto.

Flanagan, J.L. and L. Landgraf. 1968. Self-oscillating source for vocal-tract synthesizers. IEEE Trans. Audio. 16.57-64.

Fonagy, I. 1966. Electro-physiological and acoustic correlates of stress and stress perception. JSHR. 9.231-44.

Fromkin, V. 1968. Speculations on performance models. J. Linguistics 4.47-68.

Fromkin, V. and J. Ohala. 1968. Laryngeal control and a model of speech production. Working Papers in Phonetics, UCLA. 10.98-110.

Fry, D.B. 1958. Experiments in the perception of stress. L & S. 1.126-52.

Fry, D.B. 1965. The dependence of stress judgments on vowel formant structure. Proc. Vth Int'l. Cong. Phon. Sci., Münster. Basel: Karger.

Greenewalt, C.H. 1968. Bird song: Acoustics and physiology. Washington, D.C.: Smithsonian Institute.

Hadding-Koch, K. 1961. Acoustico-phonetic studies in the intonation of southern Swedish. Lund: C.W.K. Gleerup.

Hadding-Koch, K. and M. Studdert-Kennedy. 1964. An experimental study of some intonation contours. Phonetica 11.175-85.

Haggard, M. 1969. Pitch is voicing; manner is place. JAcS. 46.97.

Halliday, M.A.K. 1967. Intonation and grammar in British English. The Hague: Mouton.

Harris, K.S., T. Gay, G.N. Scholes and P. Lieberman. 1968. Some stress effects on electromyographic measures of consonant articulation. Status Report 13/14 Haskins Laboratories, New York City (Also presented at Kyoto Speech Symp. of 6th Intl. Cong. on Acoust., 29 August 1968).

Harris, Z. 1944. Simultaneous components in phonology. Language 20.181-205.

Hart, J. and A. van Katwijk. 1969. On levels of stress. IPO Annual Progress Report 4. Eindhoven, Holland.

Hirano, M. and J. Ohala. 1969. Use of hooked-wire electrodes for electromyography of the intrinsic laryngeal muscles. JSHR. 12.362-73.

Hollien, H. and J.F. Curtin. 1960. A laminagraphic study of vocal pitch. JSHR. 3.361-71.

House, A.S., and G. Fairbanks. 1953. The influence of consonant environment upon secondary acoustical characteristics of vowels. JAcS. 25.105-13.

Husson, R. 1950. Etude des phenomenes physiologiques et acoustiques fondamentaux de la voix chantée. Thésé de la Faculté de Science, Paris.

Isacenko, A.V. and H.J. Schadlich. 1963. Erzeugung künstlicher deutscher Satzintonationen mit zwei kontrastierenden Tonstufen. Monatsber. Deut. Akad. Wiss. Berlin, 6.

Jakobson, R., C.G.M. Fant, and M. Halle. 1952. Preliminaries to speech analysis. Cambridge, Mass.: M.I.T. Press.

Jassem, W. 1952. Intonation of conversational English. Travaux de la société des sciences et des lettres de Wroclaw, Series A, no. 45. Wroclaw.

Jassem, W. 1959. Phonology of Polish stress. Word 15.252-70.

Jones, D. 1909. Intonation curves. Leipzig: B.G. Teubner.

Jones, D. 1932. An outline of English phonetics, 3rd ed. New York: Dutton.

Katwijk, A. van and G.A. Govaert. 1967. Prominence as a function of the location of pitch movement. I.P.O. Annual Progress Rept. 2.115-7. Eindhoven, Holland.

Kingdon, R. 1958. The groundword of English intonation. London: Longmans, Green & Co.

Kumar, A. and K. Ojamaa. 1970. Pitch and sentence intonation. JAcS. 48.84.

Ladefoged, P. 1962. Subglottal activity during speech. Proc. IVth Int'l Cong. Phon. Sci., Helsinki. The Hague: Mouton.

Ladefoged, P. 1967. Linguistic phonetics. Working Papers in Phonetics 6. Los Angeles: UCLA Phonetics Laboratory.

Ladefoged, P. 1968. Linguistic aspects of respiratory phenonena. Annals of the New York Acad. Sci. 155.141-51.

Ladefoged, P. 1969. Three areas of experimental phonetics. London: Oxford Univ. Press.

Ladefoged, P., M.H. Draper and D. Whitteridge. 1958. Syllables and stress. Miscellanea Phonetica 3.1-14.

Ladefoged, P. and N.P. McKinney. 1963. Loudness, sound pressure and subglottal pressure in speech. JAcS. 35.454.

Lee, W.R. 1956. Fall-rise intonations in English. Eng. Stud. 37.62-72.

Lehiste, I. 1961. Some acoustic correlates of accent in Serbo-Croatian. Phonetica 7.114-47.

Lehiste, I. and G.E. Peterson. 1959. Vowel amplitude and phonemic stress in American English. JAcS. 31.428-35.

Lenneburg, E.H. 1967. Biological foundations of language. New York: Wiley.

Liberman, A.M., F.S. Cooper, D.P. Shankweiler and M. Studdert-Kennedy. 1967. Perception of the speech code. Psychol. Rev. 74.431-61.

Lieberman, P. 1960. Some acoustic correlates of word stress in American-English. JAcS. 32.451-4.

Lieberman, P. 1963. Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. JAcS. 35.344-53.

Lieberman, P. 1965. On the acoustic basis of the perception of intonation by linguists. Word 21.40-54.

Lieberman, P. 1967. Intonation, perception and language. Cambridge, Mass.: M.I.T. Press.

Lieberman, P. 1968a. Direct comparison of subglottal and esophageal pressure during speech. JAcS. 43.1157-64.

Lieberman, P. 1968b. Primate vocalizations and human linguistic ability. JAcS. 44.1574-84.

Lieberman, P. 1970. Towards a unified phonetic theory. Linguistic Inquiry 1.307-22.

Lieberman, P. and E.S. Crelin. 1971. On the speech of Neanderthal man. Linguistic Inquiry 2, No. 2.

Lieberman, P., R. Knudson and J. Mead. 1969. Determination of the rate of change of fundamental frequency with respect to subglottal air pressure during sustained phonation. JAcS. 45.1537-43.

Lieberman, P. and S.B. Michaels. 1962. Some aspects of fundamental frequency, envelope amplitude and the emotional content of speech. JAcS. 34.922-7.

Lieberman, P., M. Sawashima, K.S. Harris and T. Gay. 1970. The articulatory implementation of the breath-group and prominence: Cricothyroid muscular activity in intonation. Language 46.312-27.

Lindblom, G. 1963. On vowel reduction. Report No. 29. Stockholm: Speech Transmission Laboratory, Royal Instit. of Tech.

Lisker, L. and A.S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. Word 20.384-422.

Lisker, L., A.S. Abramson, F.S. Cooper, and M.H. Schvey. 1969. Transillumination of the larynx in running speech. JAcS. 45.1544-6.

Marler, P.F. and W.J. Hamilton. 1966. Mechanisms of animal behavior. New York: Wiley.

Matsui, E., T. Suzuki, N. Umeda and H. Omura. 1968. Synthesis of fairy tales using an analog vocal tract. Reports of the 6th International Congress on Acoustics. B.159-62. Tokyo: International Council of Scientific Unions.

Mattingly, I.G. 1966. Synthesis by rule of prosodic features. L & S. 9.1-13.

Mattingly, I.G. 1968. Synthesis by rule of general American English. Supplement to status report on speech research. New York: Haskins Laboratories.

Mead, J. and E. Agostini. 1964. Dynamics of breathing. In Handbook of physiology, respiration, I. Fenn and Rahn, eds., Washington, D.C.: American Physiological Society.

Morton, J. and W. Jassem. 1965. Acoustic correlates of stress. L & S. 8.159-81.

Müller, J. 1848. The physiology of the senses, voice and muscular motion with the mental facilities, W. Baly, translator. London: Taylor, Walton and Maberly.

Nash, R. 1967. Turkish intonation: An instrumental study. Ph.D. dissertation. Bloomington: Indiana University.

Negus, V.E. 1949. The comparative anatomy and physiology of the larynx. London: Hafner.

Ohala, J. 1970. Aspects of the control and production of speech. Working papers in phonetics 15. Los Angeles: UCLA Phonetics Laboratory.

Ohala, J. and M. Hirano. 1967. Studies of pitch changes in speech. Working papers in phonetics. 7.80-4. Los Angeles: UCLA Phonetics Laboratory.

Ohala, J. and H. Hirose. 1969. The function of the sternohyoid muscle in speech. Reports of the 1969 Autumn Meeting of the Acoustical Society of Japan. 359-60. Tokyo.

Ohala, J. and P. Ladefoged. 1970. Subglottal pressure variations and glottal frequency. JAcS. 47.104.

Öhman, S. 1968. A model of word and sentence intonation. Report 2/3. Stockholm: Speech Transmission Laboratory, Royal Instit. of Technology.

Ojamaa, K., D. Erickson and A. Kumar. 1970. Coarticulation effects on pitch. JAcS. 48.84.

Perkell, J.S. 1969. Physiology of speech production, results and implications of a quantitative cineradiographic study. Cambridge, Mass.: M.I.T. Press.

Peterson, G.E. and H.L. Barney. 1952. Control methods used in a study of the vowels. JAcS. 24.175-84.

Pike, K.L. 1945. The intonation of American English. Ann Arbor: University of Michigan.

Postal, P. 1964. Constituent structure: A study of contemporary models of syntactic description. IJAL. 30, No. 1, Part III.

Postal, P. 1968. Aspects of phonological theory. New York: Harper and Row.

Rabiner, L.R., H. Levitt and A.E. Rosenberg. 1969. Investigation of stress patterns for speech synthesis. JAcS. 45.92-101.

Revtova, L.D. 1965. The intonation of declarative sentences in current English and Russian. Presented at the 5th Int. Cong. Phon. Sci., Munster, Germany. August 1964. Summary in Phonetica 12.192.

Rigault, A. 1962. Rôle de la fréquence, de l'intensité et de la durée vocaliques dans la perception de l'accent in français. Proc. IVth Int. Cong. Phon. Sci., Helsinki. The Hague: Mouton.

Sawashima, M., A.S. Abramson, F.S. Cooper, and L. Lisker. MS. Observing laryngeal adjustments during running speech by use of a fiberoptics system. To appear in Phonetica.

Schubiger, M. 1958. English intonation, its form and function. Tübingen: Max Niemeyer.

Sonninen, A.A. 1956. The role of the external laryngeal muscles in length adjustment of the vocal cords in singing. Acta OTO Laryngol., Supp. 130.

Stetson, R.H. 1951. Motor phonetics. Amsterdam: North-Holland.

Sundberg, J. 1969. Articulatory differences between spoken and sung vowels in singers. Speech Transmission Laboratory Quarterly Status Rept. 1/1969. Stockholm: Royal Instit. Tech.

Swigart, E. and Y. Takefuta. 1968. The produced and perceived pitch of American English vowels. IEEE Trans. Audio. 16.273-7.

Timcke, R., H. von Leden, and P. Moore. 1958. Laryngeal vibrations, measurements of the glottic wave. A.M.A. Arch. Otolaryngol. 68.1-19.

Trager, G.L. and H.L. Smith. 1951. Outline of English structure, studies in linguistics, No. 3. Norman, Okla.: Battenburg.

Van den Berg, Jw. 1957. Subglottic pressure and vibrations of the vocal cords. Folia Phoniat. 2.65-71.

Van den Berg, Jw. 1958. Myoelastic-aerodynamic theory of voice production. JSHR. 1.227-44.

Van den Berg, Jw. 1960. Vocal ligaments versus registers. Current Prob. Phoniat. Logoped. 1.19-34.

Van den Berg, Jw. 1968. Sound production in isolated human larynges. Annals of the New York Acad. Sci. 155.18-27.

Vanderslice, R. 1967. Larynx versus lungs: Cricothyrometer data refuting some recent claims concerning intonation and archetypality. Working Papers in Phonetics, UCLA. 7.69-80.

Vanderslice, R. 1970. Prosodic model for orthographic-to-phonetic conversion of English. JAcS. 48.84.

Vanderslice, R. and L.S. Pierson. 1967. Prosodic features of Hawaiian English. Quar. J. Speech 53.156-66.

Wang, W. S-Y. 1962. Stress in English. Language Learning 12.69-77.

Wang, W. S-Y. 1967. Phonological features of tone. IJAL. 33.93-105.

Whitfield, I.C. 1969. Response of the auditory nervous system to simple time-dependent acoustic stimuli. Annals of New York Acad. Sci. 156.671-7.

MANUSCRIPTS FOR PUBLICATION, REPORTS, AND ORAL PAPERS*

Phonetics: An Overview. Arthur S. Abramson. In Current Trends in Linguistics, Vol. XII, ed. by Thomas A. Sebeok (The Hague: Mouton, in press).

The Perception of Speech. Michael Studdert-Kennedy. In Current Trends in Linguistics, Vol. XII, ed. by Thomas A. Sebeok (The Hague: Mouton, in press).

Physiological Aspects of Articulatory Behavior. Katherine S. Harris. In Current Trends in Linguistics, Vol. XII, ed. by Thomas A. Sebeok (The Hague: Mouton, in press).

Laryngeal Research in Experimental Phonetics. Masayuki Sawashima. In Current Trends in Linguistics, Vol. XII, ed. by Thomas A. Sebeok (The Hague: Mouton, in press).

Speech Synthesis for Phonetic and Phonological Models. Ignatius Mattingly. In Current Trends in Linguistics, Vol. XII, ed. by Thomas A. Sebeok (The Hague: Mouton, in press).

On Time and Timing in Speech. Leigh Lisker. In Current Trends in Linguistics, Vol. XII, ed. by Thomas A. Sebeok (The Hague: Mouton, in press).

A Study of Prosodic Features. Philip Lieberman. In Current Trends in Linguistics, Vol. XII, ed. by Thomas A. Sebeok (The Hague: Mouton, in press).

---

*The contents of this Report, SR 23, are included in this listing.